LETTER **Communicated by Nathaniel Daw**

Attention-Gated Reinforcement Learning of Internal Representations for Classification

Pieter R. Roelfsema
p.roelfsema@ioi.knaw.nl
Netherlands Ophthalmic Research Institute, 1105 BA Amsterdam, Netherlands, and Center for Neurogenomics and Cognitive Research, Department of Experimental Neurophysiology, Vrije Universiteit, 1081 HV Amsterdam, Netherlands

Arjen van Ooyen
Arjen.van.ooyen@falw.vu.nl
Netherlands Institute for Brain Research, 1105 AZ Amsterdam, Netherlands, and Center for Neurogenomics and Cognitive Research, Department of Experimental Neurophysiology, Vrije Universiteit, 1081 HV Amsterdam, Netherlands

Animal learning is associated with changes in the efficacy of connections between neurons. The rules that govern this plasticity can be tested in neural networks. Rules that train neural networks to map stimuli onto outputs are given by supervised learning and reinforcement learning theories. Supervised learning is efficient but biologically implausible. In contrast, reinforcement learning is biologically plausible but comparatively inefficient. It lacks a mechanism that can identify units at early processing levels that play a decisive role in the stimulus-response mapping. Here we show that this so-called credit assignment problem can be solved by a new role for attention in learning. There are two factors in our new learning scheme that determine synaptic plasticity: (1) a reinforcement signal that is homogeneous across the network and depends on the amount of reward obtained after a trial, and (2) an attentional feedback signal from the output layer that limits plasticity to those units at earlier processing levels that are crucial for the stimulus-response mapping. The new scheme is called attention-gated reinforcement learning (AGREL). We show that it is as efficient as supervised learning in classification tasks. AGREL is biologically realistic and integrates the role of feedback connections, attention effects, synaptic plasticity, and reinforcement learning signals into a coherent framework.

1 Introduction

A fundamental question in cortical neurophysiology is how neurons in sensory areas of the cortex modify their tuning to improve the animal's performance. During development, but also in adulthood, neurons in sensory

areas become tuned to features that are relevant to behavior and lose their sensitivity to features that do not carry useful information. It is still unclear how behavioral relevance influences sensory representations, and the mechanisms that guide this plasticity are only partially understood.

The question of how to change the tuning of sensory neurons has also been addressed in artificial neural networks, where it corresponds to modifying the tuning of hidden units, which are intermediate between the network's input and output layers. Learning theories have proposed various rules to change the tuning of hidden units. These theories broadly fall into three classes: unsupervised, supervised, and reinforcement learning schemes. Unsupervised learning schemes modify synapses on the basis of statistical regularities in the input (e.g., Rumelhart & Zipser, 1986; Becker & Hinton, 1992; Hinton, Dayan, Frey, & Neal, 1995; Olshausen & Field, 1996), but they do not consider the consequences of the representation for behavior. Recent neurophysiological studies have demonstrated, however, that neurons in the frontal cortex (Freedman, Riesenhuber, Poggio, & Miller, 2001), the inferotemporal cortex (Sigala & Logothetis, 2002; Baker, Behrmann, & Olson, 2002), and even the primary visual cortex (Schoups, Vogels, Qian, & Orban, 2001) become selectively tuned to variations in the input that are most relevant for behavior. This implies that unsupervised learning methods are incomplete.

In contrast, supervised learning schemes such as error backpropagation (Rumelhart, Hinton, & Williams, 1986; Bishop, 1995) change the tuning of hidden units in order to improve the network's output. These learning schemes are efficient because they form internal representations that support nonlinear input-output mappings, and they are widely applied to train artificial neural networks. However, these schemes are implausible from a neurobiological point of view, for at least two reasons (Barto, 1985; Crick, 1989). First, supervised learning schemes have to propagate specific error signals from the output layer back to the input layer (Zipser & Rumelhart, 1990). These error signals are not observed in neurophysiology. Second, a "teacher" has to specify the correct pattern of activity across the network's output layer during learning. It is unclear how a teacher could specify the target activity of all the neurons in, for example, the motor cortex.

Reinforcement learning models are much more popular in neurobiology (Barto, 1985; Montague, Dayan, Person, & Sejnowski, 1995; Schultz, Dayan, & Montague, 1997; Sutton & Barto, 1998). In neurobiologically inspired models, the output is chosen stochastically, so that the network can try various outputs for each of the input patterns. This is reminiscent of animal learning, where the animal tries out various responses until it finds the correct one. Moreover, the teacher is replaced by a global reinforcer, such as the presence or absence of reward. Biological reinforcement learning schemes modify behavior on the basis of whether the amount of reward on a particular trial is better or worse than expected. The popularity of reinforcement learning models has greatly increased in recent years, as signals predicted by

reinforcement learning theories have been found in the brain (Ljungberg, Apicella, & Schultz, 1993; Schultz et al., 1997; Schultz & Dickinson, 2000; Waelti, Dickinson, & Schultz, 2001; Schultz, 2002). Reinforcement learning has been used to train biologically inspired neural networks to perform relatively simple stimulus-response mappings (Barto, 1985; Barto & Anandan, 1985; Williams, 1992; Montague et al., 1995). However, previous reinforcement learning schemes that use biologically plausible learning rules are not as efficient as supervised learning schemes in optimizing the tuning of hidden units. The reason is that these models lack an efficient mechanism to assign credit to those hidden units that play a crucial role in the stimulus-response mapping (Barto, 1985; Williams, 1992; Bishop, 1995).

In this study, we demonstrate how this credit assignment problem can be solved in a neurophysiologically plausible way by the inclusion of an attentional signal that feeds back from the network's response selection stage to earlier processing levels. In the literature, the word *attention* is used for many different concepts. To avoid confusion, it is important to stress that this feedback signal would correspond to what psychologists call *goal-driven* (or *top-down*) selective attention.

Here we focus on tasks that are restricted in two respects. First, the tasks under study require the assignment of a unique response to each member of a set of stimuli (1-of-n deterministic categorization task). Second, the reward is delivered immediately after the network's response. Thus, we address only the "spatial" credit assignment problem of identifying units that were involved in selecting the response. We do not address the so-called temporal credit assignment problem that arises when rewards are delivered after a delay or when the animal progresses through a sequence of stages before reward delivery. This temporal credit assignment problem has been the focus of a number of previous studies (Sutton & Barto, 1998; Montague et al., 1995).

To gain insight into the spatial credit assignment problem, we start from the observation that areas of the cerebral cortex interact with each other through a dense network of feedforward and feedback connections (Felleman & Van Essen, 1991). Feedforward connections map sensory stimuli onto motor responses. They propagate neuronal activity from sensory cortex to association cortex and from there to the motor cortex. However, there are also feedback connections, which propagate activity from the motor cortex back to the sensory cortex. Feedback connections mediate goal-driven attentional effects (Desimone & Duncan, 1995; Moore, 1999; Treue & Martínez Trujillo, 1999), and attention is necessary for neuronal plasticity at early processing levels (Ahissar & Hochstein, 1993; Schoups et al., 2001). The precise role of attention in learning, however, is not well understood.

In this study, we therefore investigate the consequences of attentional feedback for learning in a neural network. We use two factors to modulate synaptic plasticity. The first factor, δ, encodes whether the amount of reward obtained after a trial is better or worse than expected (e.g., Montague

et al., 1995). This δ has been used in many previous studies on reinforcement learning. It is a global signal that is delivered to all units regardless of whether they were involved in the network's choice. The second factor is the attentional feedback signal from units in the output layer. This factor gates the plasticity of units at earlier processing levels responsible for the network's output. This credit assignment signal does not depend on reward and is the distinguishing feature of the new learning scheme. We call the new scheme AGREL, which stands for *attention-gated reinforcement learning*. AGREL provides a new theoretical link between supervised learning and biologically inspired reinforcement learning theories. We will demonstrate that AGREL is as powerful as previous supervised learning schemes in deterministic categorization tasks and yet plausible from a neurophysiological point of view.

2 Task and Network Design

We will use a neural network to simulate the selection of behavioral responses by an animal that is learning to classify stimuli into a number of categories. In a real experiment, the animal would be required to associate a unique action with every stimulus category (see, e.g., Schoups et al., 2001; Freedman et al., 2001; Sigala & Logothetis, 2002; Baker et al., 2002). We will use P stimuli that have to be categorized into C mutually exclusive classes. In the neural network, a stimulus is presented onto the input layer on every trial (see Figure 1). Activity is then propagated to a hidden layer (a higher sensory area), and from there to the output layer (representing the motor cortex) that has an output unit for every category. If presented with stimulus p, the task of the network is to respond with a target pattern $\mathbf{t}^p$ in which all output units k have activity 0, except unit $k = c_p$, which encodes the target category and has activity 1. If the network chooses the correct class, it receives a reward. If not, it receives nothing. Thus, in case of an error, the network is not informed about the class that should have been chosen if there are more than two categories. On these trials, the feedback from the environment is less informative than that given by supervised learning schemes. After the trial, synaptic connections are updated. In a neurobiological model, connections should be changed on the basis of information that is available locally, at the synapse. We will show that AGREL can form useful internal representations, even if this neurobiological constraint is taken into account. We emphasize that AGREL is not an improved method for the training of artifical neural networks.

The new model is closely related to error backpropagation (BP), an efficient supervised learning method for training neural networks in nonlinear classification tasks (Rumelhart et al., 1986; Bishop, 1995). In the sequel, we will compare AGREL to BP, and we will therefore first outline how error BP is used to train neural networks for classification. This will allow us to

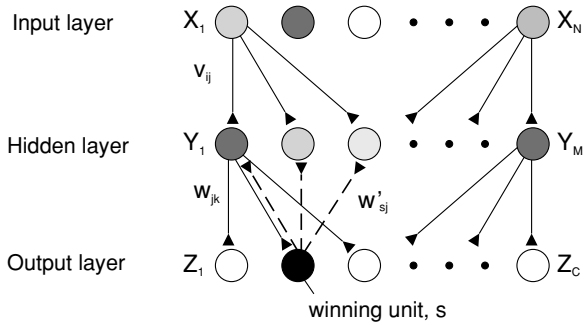

Figure 1: Three-layer neural network that is trained to perform a classification task. The task of the network is to activate a single output unit c_p that encodes the class of the stimulus pattern p. There are N units in the input layer, M in the second (hidden) layer, and C in the output layer. Connections v_{ij} propagate activity from the input layer to the hidden layer, and connections w_{jk} in turn propagate the activity from the hidden to the output layer. The winning output unit, s, feeds its activity back to the hidden layer through connections w'_{sj} (dashed lines).

indicate the neurobiologically implausible features of BP (see also Barto, 1985; Crick, 1989).

3 Error Backpropagation

We first explain how BP is used to train the three-layer network of Figure 1 in a classification task. On each trial, an input pattern p is presented to the network's input layer with N units (with activities X_i^p) and activity is propagated to the hidden layer with M hidden units through connections v_{ij}. The activity of units in this layer, Y_j^p, depends on their input according to a nonlinear activation function. Here we use the logistic function (but we note that our results do not depend critically on this choice),

$$Y_j^p = \frac{1}{1+\exp\left(-h_j^p\right)} \quad \text{with} \quad h_j^p = \sum_{i=0}^{N} v_{ij} X_i^p. \tag{3.1}$$

Hidden units j also have a bias weight, v_{0j}. Connections w_{jk} in turn propagate the activity from the hidden to the output layer. If the task is classification, then it is advantageous to choose the softmax activation function for the output units (Bishop, 1995),

$$Z_k^p = \frac{\exp\left(a_k^p\right)}{\sum_{k'=1}^{C} \exp\left(a_{k'}^p\right)} \quad \text{with} \quad a_k^p = \sum_{j=0}^{M} w_{jk} Y_j^p. \tag{3.2}$$

Output units k also have a bias weight, w_{0k}. After each trial, error BP determines the error at each output unit, which is related to the difference between the actual and target output of the unit. A suitable definition for the total error is given by the cross-entropy (van Ooyen & Nienhuis, 1992; Bishop, 1995),

$$Q_p = -\sum_{k=1}^{C} t_k^p \ln Z_k^p. \tag{3.3}$$

Here, $t_{c_p}^p = 1$, and $t_k^p = 0$ for $k \neq c_p$ (the components of target pattern $\mathbf{t}^p$). To reduce the error in the output, error BP changes connection weights along the error gradient. Weights between the hidden and output layer are changed according to

$$\Delta w_{jk} = -\beta \frac{\partial Q_p}{\partial w_{jk}} = \beta Y_j^p \left(t_k^p - Z_k^p\right), \tag{3.4}$$

and weights between the input and hidden layer according to

$$\Delta v_{ij} = -\beta \frac{\partial Q_p}{\partial v_{ij}} = \beta X_i^p Y_j^p \left(1 - Y_j^p\right) \sum_{k=1}^{C} \left(t_k^p - Z_k^p\right) w_{jk}, \tag{3.5}$$

where β is a parameter that determines the learning rate. These equations are biologically implausible, because they imply that a "teacher" has to specify t_k^p after each trial for each of the output units. It is unlikely that a teacher in the brain could specify the target activity of all output neurons in, for example, the frontal or motor cortex. Furthermore, the change in synaptic weight Δv_{ij} in equation 3.5 depends on a sum over error signals $(t_k^p - Z_k^p)$ at the output layer multiplied by the connection strengths w_{jk}. Thus, each hidden unit must know its own individualized error signal (see also Barto, 1985). A system that calculates these error signals does not appear to exist in the brain. Error-related signals that are found in neurophysiology are global and not fine-tuned to individual neurons (Montague et al., 1995; Schultz et al., 1997; Schultz & Dickinson, 2000).

4 AGREL: Attention-Gated Reinforcement Learning

Reinforcement learning schemes do not require a teacher that reveals the target output because the only information required for learning is the delivery or omission of a reward (Barto, 1985; Barto & Anandan, 1985; Sutton & Barto, 1998). Based on this information, the system computes a global error signal, δ, which reflects changes in reward expectancy. Recent neurophysiological studies demonstrated that such a signal is indeed computed,

by dopamine neurons of the midbrain (Montague et al., 1995; Schultz et al., 1997; Schultz & Dickinson, 2000). The dopamine neurons carry information about rewards expected on the current trial, even if reward delivery is somewhat delayed. Here, we will study AGREL only in tasks where rewards are delivered immediately after correct responses. In that case, the reward-evoked dopamine activity (as well as δ) reflects the difference between the amount of reward that was expected before the response and the amount received afterward (Fiorillo, Tobler & Schultz, 2003; Morris, Arkadir, Nevet, Vaadia & Bergman, 2004). A disadvantage of previous biological reinforcement learning schemes is that they are not as efficient as BP in optimizing the tuning of units in the hidden layer. There is no mechanism that identifies the hidden units that are responsible for the outcome of a trial. AGREL solves this credit assignment problem by including cortical feedback connections that mediate attentional effects. Thus, AGREL combines two signals that jointly determine plasticity: (1) the global error signal δ and (2) an attentional signal that feeds back from the output layer to previous layers.

When a new input pattern is presented, activity is propagated to the output layer, just as in BP. A crucial feature of AGREL is that units in the output layer engage in a competition. On every trial, one output unit wins and gets activity 1, while the other output units get activity 0. AGREL uses the stochastic softmax rule to determine the probability of choosing unit k as the winning unit,

$$\Pr\left(Z_k^p = 1\right) = \frac{\exp\left(a_k^p\right)}{\sum_{k'=1}^{C} \exp\left(a_{k'}^p\right)} \quad \text{with} \quad a_k^p = \sum_{j=0}^{M} w_{jk} Y_j^p. \tag{4.1}$$

Equation 4.1 is similar to equation 3.2, but it now describes the probabilities in a competitive selection process. An increase in the synaptic input a_k^p to output unit k enhances the probability of selecting this unit and decreases the probability of selecting others.

If the network chooses the correct output unit for a particular stimulus, it receives a reward r. We will assume that r equals 1. No reward is given in case of misclassification. After each trial, the synaptic weights are updated according to a simple and physiologically plausible Hebbian rule, which states that the change in synaptic weight depends on the product of pre- and postsynaptic activity (Malinow & Miller, 1986; Gustafsson & Wigstrom, 1988). For the weights between the hidden units and output units, the factor δ modulates the Hebbian plasticity,

$$\Delta w_{jk} = \beta Y_j^p Z_k^p f(\delta). \tag{4.2}$$

Note that this equation implies that only connections onto the winning output unit (with activity $Z_k^p = 1$) are changed, because the other output

units have activity 0 after the competition. On rewarded trials, δ equals the difference between the amount of reward that was obtained and the amount that was expected for a particular stimulus,

$$\delta = r - E^p(r). \tag{4.3}$$

$E^p(r)$ is the average amount of reward that is expected with stimulus p, and this equals $\Pr(Z_{c_p}^p = 1)$, the probability that the correct output unit is chosen. On rewarded trials, δ therefore equals $1 - \Pr(Z_{c_p}^p = 1)$. The probability $\Pr(Z_{c_p}^p = 1)$ can be determined by evaluating activity in the output layer at the start of the competition (an alternative method to compute δ is described in section 7.2). On unrewarded trials, however, plasticity in AGREL does not depend on $E^p(r)$, and δ is set to -1.

Unexpected rewards are especially valuable in learning. In AGREL, δ therefore influences synaptic plasticity through an expansive function $f(\delta)$ that can be implemented at the synapse. $f(\delta)$ takes large values if δ is close to 1, that is, when actions are rewarded unexpectedly:

$$f(\delta) = \begin{cases} \delta/(1-\delta); & \delta \geq 0 \\ \delta; & \delta = -1 \end{cases}. \tag{4.4}$$

The weights v_{ij} between the input layer and the hidden layer are also modified according to a Hebbian rule that depends on $f(\delta)$. However, here a second factor, $fb_{Y_j}^p$, which equals the feedback arriving from the output layer at unit Y_j, also influences plasticity:

$$\Delta v_{ij} = \beta X_i^p Y_j^p f(\delta) fb_{Y_j}^p \quad \text{with} \quad fb_{Y_j}^p = \left(1 - Y_j^p\right) \sum_{k=1}^{C} Z_k^p w'_{kj}. \tag{4.5}$$

After the competition, only the winning output unit has activity 1, and the other output units have activity 0. Equation 4.5 therefore reduces to

$$\Delta v_{ij} = \beta X_i^p Y_j^p f(\delta)\left[w'_{sj}\left(1 - Y_j^p\right)\right], \tag{4.6}$$

where w'_{sj} stands for feedback of the winning unit s (dashed connections in Figure 1). This feedback signal gates the plasticity of connections v_{ij} from the input layer to hidden unit j. The factor $(1 - Y_j^p)$ reduces the effect of feedback on the plasticity of highly active units. Note that synapses can implement equations 4.2, 4.4, and 4.6, because the required information is available locally.

Cortical anatomy and neurophysiology suggests that feedforward and feedback connections are reciprocal (Felleman & Van Essen, 1991; Salin & Bullier, 1995). In AGREL, the plasticity of feedback connections w'_{kj} is

therefore also governed by equation 4.2, so that the strength of feedforward and feedback connections becomes proportional during training (deviations from exact reciprocity are investigated in section 5.4). The consequence of this reciprocity is that hidden units that provide most excitation to the winning output unit also receive strongest feedback. Feedback thereby assigns credit to hidden units that are responsible for the choice of action.

4.1 Average Weight Changes in AGREL. We now compute the average weight changes in AGREL. First, we compute the average change in synaptic weights w_{jk} between the hidden and output layer. Note that as a result of equation 4.2, only connections to the winning output unit are updated, since for the other units, $Z_k^p = 0$. The correct output unit $k = c_p$ is selected with probability $\Pr(Z_{c_p}^p = 1)$, causing an average change in weight w_{jc_p} across trials of

$$E\left(\Delta w_{jc_p}\right) = \Pr\left(Z_{c_p}^p = 1\right)\beta Y_j^p \delta/(1-\delta) = \beta Y_j^p\left[1 - \Pr\left(Z_{c_p}^p = 1\right)\right]. \quad (4.7)$$

An erroneous output unit $k \neq c_p$ is selected with probability $\Pr(Z_k^p = 1)$, and the average change in weights w_{jk} equals

$$E(\Delta w_{jk}) = \Pr\left(Z_k^p = 1\right)\beta Y_j^p f(\delta) = -\beta Y_j^p \Pr\left(Z_k^p = 1\right); \quad k \neq c_p. \quad (4.8)$$

Combining equations 4.7 and 4.8 yields

$$E(\Delta w_{jk}) = \beta Y_j^p\left[t_k^p - \Pr\left(Z_k^p = 1\right)\right]. \quad (4.9)$$

We can compute the average change in weights v_{ij} across trials from equation 4.6:

$$\begin{aligned} E(\Delta v_{ij}) &= \sum_{s=1}^{C} \Pr\left(Z_s^p = 1\right)\beta X_i^p Y_j^p f(\delta)\left(1 - Y_j^p\right) w'_{sj} \\ &= \Pr\left(Z_{c_p}^p = 1\right)\beta X_i^p Y_j^p\left(1 - Y_j^p\right)\frac{\delta}{1-\delta} w'_{c_p j} \\ &\quad - \sum_{k \neq c_p} \Pr\left(Z_k^p = 1\right)\beta X_i^p Y_j^p\left(1 - Y_j^p\right) w'_{kj} \\ &= \beta X_i^p Y_j^p\left(1 - Y_j^p\right)\sum_{k=1}^{C}\left(t_k^p - \Pr\left(Z_k^p = 1\right)\right) w'_{kj}. \end{aligned} \quad (4.10)$$

A comparison of equations 4.9 and 4.10 to equations 3.4 and 3.5 points to the central result of this study: weight changes in AGREL are, on average, the same as those in BP. This implies that AGREL can solve all 1-of-n classification tasks that can be solved by BP. This is a remarkable result since

there is no teacher, and correct classification can be learned only by trial and error.

4.2 Analysis of the Variance of Weight Changes in AGREL. The changes in synaptic weights in AGREL are, on average, the same as in error BP. However, in AGREL, they depend on the stochastic competition between units in the output layer. It is therefore of interest to compute the variance of the changes in synaptic weights across trials. The variance in the weight change Δw_{jc_p} of connections to the correct output unit equals

$$\begin{aligned}\mathrm{Var}\left(\Delta w_{jc_p}\right) &= E\left\{\left(\Delta w_{jc_p}\right)^2\right\} - \left\{E\left(\Delta w_{jc_p}\right)\right\}^2 \\ &= \left(\beta Y_j^p\right)^2 \frac{\left(1 - \Pr\left(Z_{c_p}^p = 1\right)\right)^3}{\Pr\left(Z_{c_p}^p = 1\right)}, \end{aligned} \tag{4.11}$$

and the variance in the weight change Δw_{jk} of connections to the other output units ($k \neq c_p$) equals

$$\mathrm{Var}(\Delta w_{jk}) = \left(\beta Y_j^p\right)^2 \Pr\left(Z_k^p = 1\right)\left(1 - \Pr\left(Z_k^p = 1\right)\right). \tag{4.12}$$

A similar computation yields the variance in the change of weights between the input and hidden layer:

$$\begin{aligned}\mathrm{Var}(\Delta v_{ij}) = \left(\beta X_i^p Y_j^p \left(1 - Y_j^p\right)\right)^2 \Bigg\{ & \frac{\left(1 - \left(\Pr\left(Z_{c_p}^p = 1\right)\right)^2\right.}{\Pr\left(Z_{c_p}^p = 1\right)} w'_{c_p j} \\ & + \sum_{k \neq c_p} \Pr\left(Z_k^p = 1\right) {w'_{kj}}^2 - \left(\sum_{k=1}^{C} \left[t_k^p - \Pr\left(Z_k^p = 1\right)\right] w'_{kj}\right)^2 \Bigg\}. \end{aligned} \tag{4.13}$$

We note that $\Pr(Z_{c_p}^p = 1)$ appears in a denominator in equations 4.11 and 4.13, and also that the variance increases with the square of the learning rate, β^2. Thus, the variance caused by the stochastic choices is high for large values of $\beta^2 / \Pr(Z_{c_p}^p = 1)$, that is, if the learning rate is high and the probability of correct classification is small. This variance is not homogenous across weight space, but is highest in regions with a small $\Pr(Z_{c_p}^p = 1)$. It can be seen, however, that AGREL tends to escape from these high-variance regions. The repeated choice of incorrect actions $k \neq c_p$ reduces the probability that these actions are chosen again, since AGREL weakens the connections to the corresponding output units (see equation 4.8). The decrease in the probability of choosing incorrect outputs causes a continuous increase in $\Pr(Z_{c_p}^p = 1)$,

since actions are chosen in a competitive manner. Therefore, AGREL tends to leave the regions of weight space that are associated with a high variance.

5 Benchmark Problems

To compare the efficiency of AGREL to that of BP, we carried out a number of neural network simulations. The general layout of the simulated neural networks was as in Figure 1. $f(\delta)$ of equation 4.4 increases rapidly for values of δ very close to 1. In the simulations, $f(\delta)$ was therefore clipped at $50/\beta$ if it reached a value that was larger. We first compare AGREL to the error BP algorithm on two nonlinear classification tasks that can be solved by small networks. Then we investigate whether AGREL can be used to train a larger network on a more difficult problem.

5.1 Exclusive-or. The exclusive-or (XOR) problem is a classical nonlinear classification task. There are two input units and four input patterns: 00 (both input units off), 01, 10, and 11 (both on). There are two output units, one of which should be active for input patterns 00 and 11, and the other for 01 and 10. We compared AGREL to BP by measuring the median number of iterations required to reach a performance criterion. The criterion used for AGREL was that the probability of correct classification was at least 75% for each of the input patterns. The criterion for BP was that the difference between the activity of units in the output layer and their target values was less than 0.25 for all output units and for each of the input patterns. Table 1 shows for both models the median number of iterations (presentations of the entire stimulus set; one iteration equals four trials) required to reach criterion. Initial connection weights were drawn from a uniform distribution in the interval $[-0.25, 0.25]$, and the optimal value was determined for the learning rate β. We tested one model with two and one with three hidden

Table 1: Comparison of the Speed of Convergence of AGREL and Standard Error Backpropagation.

	BP		AGREL	
	Iterations	β	Iterations	β
XOR (2 hidden units)	366	0.6	535	0.35
XOR (3 hidden units)	218	0.9	474	0.45
Counting 2 inputs	33	2.0	157	0.4
Counting 3 inputs	71	1.5	494	0.25
Counting 4 inputs	126	1.0	1316	0.1
Mine detection	120	0.45	492	0.05

Notes: Indicated is the median number of iterations until criterion performance was reached. βs are optimal learning rates that yielded fastest convergence.

units. The model with two hidden units did not converge within 25.000 iterations in some of the runs (2 out of 10 for AGREL and 2 out of 10 for BP). In the other runs, the median number of iterations required by BP and AGREL were 366 and 535, respectively. The model with three hidden units converged in all runs with sufficiently small β and required fewer iterations: 218 for BP and 474 for AGREL.

5.2 Counting. In the counting task, there are N input units (here we used $N = 2, 3$, or 4) and the network has to determine the number of input units that are "on". There are $N + 1$ output units (the classes are 0, 1, . . . , N), and we used $N + 1$ hidden units. The first output unit should be active if all the input units are off, the second if one of the input units is on, the third if two are on, and so on. In a network with two input units, AGREL required a median of 157 iterations, about five times the number required by BP (see Table 1). In the case of three input units, AGREL reached criterion after 494 iterations, which is 7 times as many as BP, and with four input units AGREL required 1316 iterations, 10 times as many BP. Thus, AGREL converges more slowly than BP for this problem, especially if N is large. This can be explained: AGREL has to try various output units for each of the input patterns before it can determine the correct one, which slows the convergence, especially if there are many classes. This effect does not occur in BP, since the "teacher" indicates the correct class on every trial. Therefore, the number of iterations required increases faster with the number of classes for AGREL than for BP.

5.3 Mine Detection. To test AGREL on a more complex benchmark problem, we used the mine detection task of Gorman and Sejnowski (1988). The database for this task can be downloaded from various websites. The task is to classify sonar returns from undersea targets as rocks or mines. There are 208 input patterns, 111 mines, and 97 rocks. The sonar data are presented as a vector across the 60 input units of the network. When we used a network with 12 hidden units, AGREL reached criterion after a median of 492 iterations (see Table 1). BP converged after a median of 120 iterations, which is in the same range as the results of Gorman and Sejnowski (1988) and about four times as fast as AGREL.

5.4 Reciprocity of Feedforward and Feedback Connections. So far, we have assumed that the strength of the feedback connections w'_{kj} is the same as the strength of the corresponding feedforward connections w_{jk}. It is not clear whether such a precise equivalence of feedforward and feedback connections in the cortex is enforced by development and whether it would hold at the start of training. We therefore ran additional simulations where the initial strengths of feedforward connections and feedback connections were chosen independently, from uniform distributions in the interval $[-0.25, 0.25]$. If w_{jk} and w'_{kj} are changed in the same way during training, their

strengths tend to become similar. In biology, the weight changes Δw_{jk} and $\Delta w'_{kj}$ may not be exactly the same, however, and we therefore also included a noise term in the updating of feedback connections:

$$\Delta w_{jk} = \beta Y_j^p Z_k^p f(\delta) \tag{5.1}$$

$$\Delta w'_{kj} = \beta(1+\eta) Y_j^p Z_k^p f(\delta). \tag{5.2}$$

Here η is a gaussian noise term with a mean of 0 and a standard deviation of 0.2. With this modified scheme, AGREL required a median of 359 iterations to solve the XOR problem with three hidden units. This is similar to the number of iterations required by AGREL if feedforward and feedback connections were identical (see Table 1). We conclude that feedforward and feedback connections need not be identical at the start of training and also that some noise in the updating of connections does not deteriorate learning.

5.5 Summary of Benchmark Results. AGREL converges somewhat more slowly than BP. In our task set, the ratio between the number of iterations required by AGREL and BP varied between 1.5 and 10 (see Table 1). If there are many categories, the random sampling of output units increases stochasticity, which decreases the convergence rate (and therefore the optimal value for β). We note, however, that this is inevitable for any reinforcement learning algorithm, as it has to find the correct category by trial and error. AGREL tolerates small differences in the strength of corresponding feedforward and feedback connections and does not require them to be identical at the start of training. We conclude that AGREL fares well on both small and large benchmark problems.

6 Changes in the Tuning of Sensory Neurons Due to Training in Categorization Tasks

The benchmark tests indicate that AGREL can be used to train artificial neural networks in a wide range of classification tasks. It should be emphasized, however, that AGREL is actually designed as a model for the neurophysiology of learning. We therefore now investigate whether AGREL can account for changes in the tuning of neurons in sensory areas that are induced by categorization training.

6.1 Face Categorization Task. Three recent studies investigated the effect of categorization training on the tuning of neurons in the inferotemporal cortex (Baker et al., 2002; Sigala & Logothetis, 2002; Freedman, Riesenhuber, Poggio, & Miller, 2003), which is a region of visual cortex that is involved in object recognition (Tanaka, 1995). In the study by Sigala and Logothetis (2002), monkeys were trained to classify face stimuli. The animals had to categorize 10 line drawings of faces into two classes. Each face consisted of

an outline and four features that varied between stimuli: eye separation, eye height, mouth height, and nose length (see Figure 2A). Each of these features could take three values. Two of the features, eye separation and eye height, were called diagnostic, as they allowed separation between classes along a linear category boundary (see Figure 2B). The stimuli were not linearly separable by using the other two, nondiagnostic features, mouth height and nose length. On each trial, the monkeys saw one stimulus and then pressed one of two levers to indicate the category. Thereafter, the animals received a reward, but only if they chose the correct category.

The animals required more than 2000 trials to learn the categorization task (N. Sigala, personal communication, January 2003). After completion of training, single neurons were recorded in the inferotemporal cortex, and their tuning to diagnostic and nondiagnostic features was compared. Strength of tuning for a particular feature dimension was quantified with the selectivity index (SI), defined as

$$SI = (R_{\max} - R_{\min})/(R_{\max} + R_{\min}), \tag{6.1}$$

where $R_{\max}$ is the response strength evoked by the best feature value and $R_{\min}$ the response to worst feature value on this dimension. Figure 2C shows the average selectivity index for diagnostic and nondiagnostic features for 96 inferotemporal neurons. Most neurons had stronger tuning to the diagnostic than to the nondiagnostic features. This result is remarkable, since it shows that neurons in this visual area become tuned to feature variations that are most relevant to behavior. To investigate whether AGREL can account for these changes in neuronal tuning, we used a model with four input units, four hidden units, and two output units (one for each category). Each input unit encoded one of the four features and had activity 0, 0.5, or 1 (see Figure 2B). We used a learning rate (β) of 0.1, and initial synaptic weights drawn from a uniform distribution in the interval $[-1.25, 1.25]$. With these parameters, it took an average of 630 trials (S.D. = 120, $N = 24$ simulations) before the probability of correctly classification for every pattern was larger than 75%. Thus, the model learned substantially faster than the monkeys. We also investigated how the selectivity index changed as a result of training. The initial selectivity index before the start of training was determined by the random pattern of synaptic weights, and it therefore did not differ between diagnostic and nondiagnostic features (see the square in Figure 2D). Figure 2D shows the selectivity for diagnostic and nondiagnostic features for 96 hidden units after training. Note that most points lie above the diagonal, which indicates that hidden units became more selective for diagnostic than for nondiagnostic features ($p < 10^{-10}$, paired t-test), with an average selectivity index of 0.27 for diagnostic and 0.17 for nondiagnostic features. This indicates that AGREL can explain why categorization training induces a selective representation of features in sensory

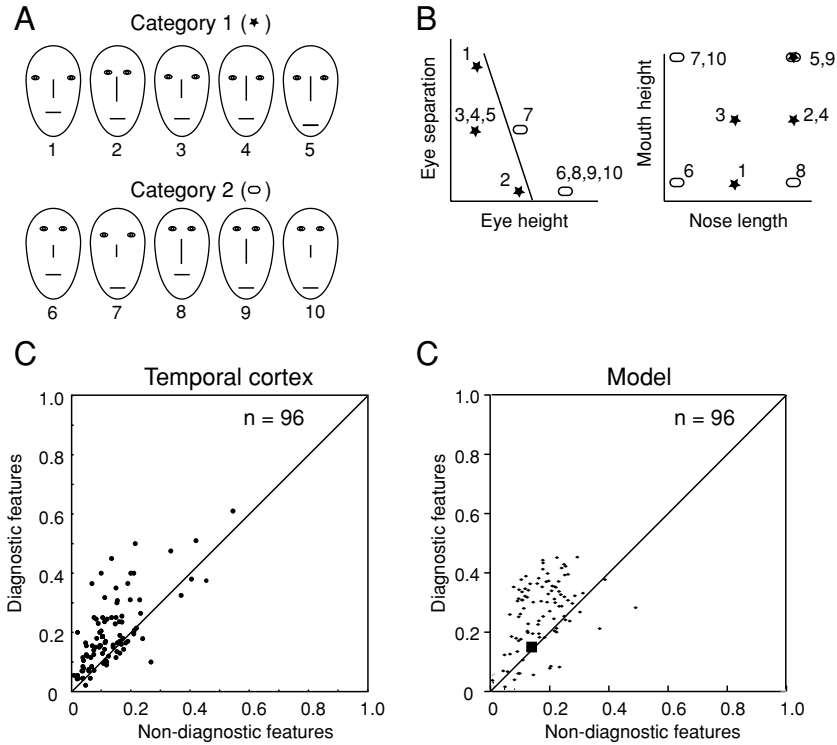

Figure 2: Representation of diagnostic and nondiagnostic features after training in a face categorization task. (A) The task is to classify faces into two categories. (B) (Left) Eye separation and eye height are diagnostic features that allow correct classification along a linear category boundary (straight line). (Right) Mouth height and nose length are nondiagnostic features that do not allow correct classification along a linear category boundary. (C, D) Selectivity indices of neurons in the inferotemporal cortex (C) and of hidden units in the model (D) for diagnostic and nondiagnostic features, after training in the categorization task. If the neuronal response strength differentiates between feature values, then the selectivity index is high. (Abscissa) Average selectivity index for the two nondiagnostic features. (Ordinate) Average selectivity index for the two diagnostic features. Most inferotemporal neurons and hidden units are better tuned to diagnostic features than to nondiagnostic features. The square in *D* shows average selectivity index of hidden units before the start of training, caused by random initialization of synaptic weights between the input and hidden layer. The results are pooled across 24 simulations, in a network with four hidden units.

areas that support classification and are therefore most useful for the task at hand.

6.2 Orientation Discrimination Task. Categorization training can also influence the representation of a single feature. Neurons in the frontal cortex (Freedman et al., 2001) and the primary visual cortex (Schoups et al., 2001) become most sensitive to feature variations close to the boundary between two categories. Here we will focus on the results of Schoups et al. (2001), who investigated the effect of orientation discrimination training on the orientation tuning of neurons in the primary visual cortex of monkeys. The animals were trained to discriminate small differences in the orientation in one grating (see the lower grating in Figure 3A), while ignoring another grating (see the upper grating in Figure 3A). At the start of training, the monkeys were able to discriminate reliably only between orientations that differed by more than 15 degrees. The monkeys were trained for tens of thousands of trials in the orientation discrimination task, and their orientation discrimination thresholds gradually decreased to values between 0.5 and 2 degrees. The advantage of this design is that neurons with receptive fields at the upper and lower grating were activated equally often during training and with the same visual stimuli. Nevertheless, only neurons that were activated by the lower grating conveyed information relevant to behavior, whereas neurons activated by the upper grating location did not.

After this training phase, recordings were made from neurons in the primary visual cortex with receptive fields at the lower or upper grating. For each neuron, the tuning for orientation was determined while the monkeys were only required to look at the fixation point; they were not engaged in the orientation discrimination task. Some of these orientation tuning curves are reproduced in Figure 3B. Cells 1 and 5 had preferred orientations differing substantially from the trained orientation, and changes in grating orientation around the trained orientation hardly influenced their response. Thus, these cells did not convey information that was relevant for solving the task. For cell 3, the preferred orientation was exactly at the trained orientation. Again, small changes in grating orientation around the trained orientation did not strongly influence the neuron's firing rate. The situation is different for cells 2 and 4, which had a preferred orientation that differed from the trained orientation by about 15 degrees. The trained orientation is at the steepest part of their tuning curve, and changes in grating orientation around the trained orientation had a strong effect on the response strength. Thus, these neurons convey most of the information required to solve the task (see also Vogels & Orban, 1990).

To quantify the effect of training across the population of neurons, the average slope of the tuning curve at the trained orientation was determined for groups of neurons with different preferred orientations (see the thick line segments in Figure 3B). The comparison of interest is between the slopes of tuning curve of neurons with receptive fields at the trained and nontrained

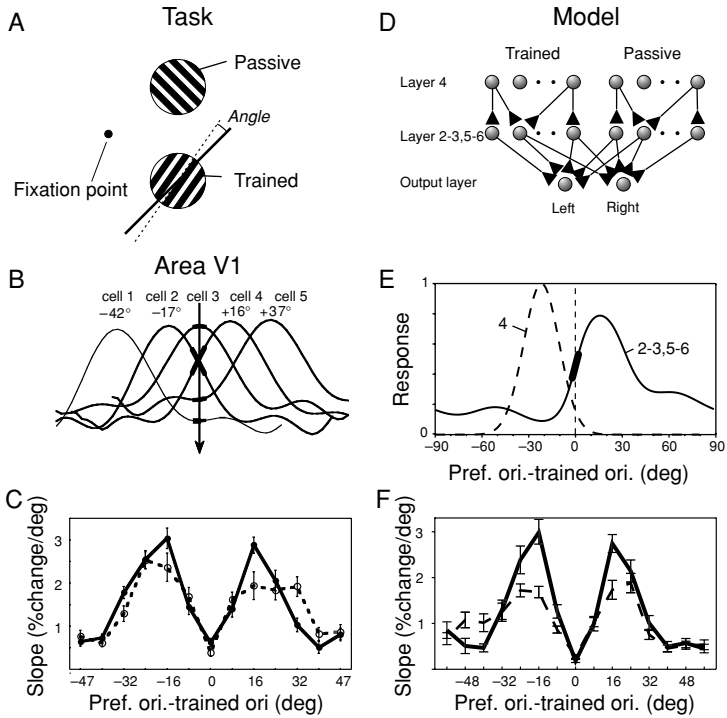

Figure 3: Effects of training in an orientation discrimination task on orientation tuning in area V1. (A) Monkeys had to decide whether the orientation of the lower grating was leftward or rightward tilted from the right oblique orientation (continuous line, 45 degrees). The upper grating could be ignored. (B) Orientation tuning curves of neurons in area V1. Arrow, trained orientation. Thick line segments indicate the slope of the tuning curves at the trained orientation. (C) The slope of tuning curves at the trained orientation as a function of the neurons' preferred orientation. The slope is highest for trained neurons (thick curve) with a preferred orientation that differs from the trained orientation (TO) by 12 to 20 degrees. The slope of the tuning curve of nontrained neurons (dashed curve) is shallower. (D) The neural network consisted of input units with gaussian orientation tuning (corresponding to cortical layer 4) and hidden units (cortical layers 2, 3, 5, and 6) at a trained and a passive retinotopic location. There were two output units; one had to be active in response to gratings at the trained location that were tilted to the left and the other to gratings tilted to the right. Orientations presented at the passive location contained no information for classification. (E) Example tuning curve of a unit in the input (dashed curve) and hidden layer (continuous curve) after training to categorize orientations that differed by 2 degrees. (F) The slope of tuning curves of hidden units as a function of preferred orientation. The slope is higher for units that respond to the trained location (continuous curve) than for units that are stimulated passively (dashed curve).

grating location. It can be seen that the tuning curves of neurons that responded to the trained grating had a steeper slope at the trained orientation than the tuning curves of neurons that responded to the irrelevant grating (see Figure 3C). Remarkably, this steepening of tuning curves is observed only for the more useful neurons with preferred orientations that differed from the trained orientation by 12 to 20 degrees. This sharpening of tuning curves was observed only in supra- and infragranular layers, but not in layer 4, which is the input layer of the primary visual cortex (Schoups et al., 2001). These results are in accordance with other results that categorization training induces a sharper tuning at the boundary between categories (Freedman et al., 2001). A spectacular aspect of this study is that these changes can be observed even in the primary visual cortex, the lowest area in the visual cortical processing hierarchy.

To investigate whether AGREL can account for these changes in neuronal tuning, we used a model with an input layer that corresponds to layer 4 of area V1, with 20 input units at two retinotopic locations (see Figure 3D). The input units had gaussian tuning curves with a σ of 12 degrees, and a half-width at half-height of 15 degrees (see Figure 3E), which is well within the biologically plausible range (Vogels & Orban, 1990). The preferred orientations of the input units differed in steps of 9 degrees. There were also 64 hidden units at each retinotopic location. The learning rate β was set to 0.02. To start with a sufficient variety of tuning curves in the hidden layer, the model was first trained with an output layer with 12 units to categorize gratings with 12 different orientations (0, 15, . . . , 165 degrees) that could appear at either location. Training was stopped when performance was at least 75% correct for each orientation at both locations, which occurred after an average of 697 presentations of the stimulus set. This initial training phase corresponds to the visual experience of the monkeys prior to the orientation discrimination task. Then a model was used with an output layer consisting of two units, and newly initialized connections between the hidden layer and the output layer, while the connections between the input layer and the hidden layer remained the same. This model was trained to categorize two orientations that differed by 2 degrees presented at one location, while an irrelevant orientation was presented at the other, passive location. This second training phase took on average 3500 trials (S.D. = 1300 trials, $N = 10$ simulations). Again, the model improved more rapidly than the monkeys, which needed at least 10 times more trials during training. Training induced a steepening of the slope of tuning curves in the hidden layer at the trained orientation (see Figure 3F). This occurred only at the trained location and only for hidden units with a preferred orientation that differed by about 15 degrees from the trained orientation.

The similarity between model and experiment (see Figures 3C and 3F) is remarkable, as this close match was achieved by varying only two parameters of the model: the amount of training in the initial phase and the sharpness of tuning of the input units. Without the initial orientation categorization training, the tuning curves in the hidden layer are determined

by the random initialization of synaptic weights from the input layer. In this case, the slopes of the tuning curves of the hidden units are relatively small, which causes a general downward shift of the two curves of Figure 3F (data not shown). Variations in the width of the tuning curves in the input layer influence the degree to which hidden units that prefer orientations differing from the trained orientation are affected by the training. If tuning curves in the input layer are narrow, the increase in slope at the trained orientation is restricted to units that prefer orientations close to the trained orientation (but not precisely at this orientation). Training with broad tuning curves in the input layer also increases the slope of the tuning curve of units that have a preferred orientation that is further from the trained orientation.

These simulation results, taken together, indicate that AGREL can explain how categorization training increases the sensitivity of sensory neurons to feature variations that are relevant for the task at hand. AGREL causes a selective representation of feature dimensions that support classification (see Figure 2) and sharpens neuronal tuning at the boundary between categories (see Figure 3).

7 Extensions of AGREL

AGREL uses two factors to gate synaptic plasticity: the global factor δ and the attentional feedback signal. It is possible to define many different learning schemes on the basis of the same principle. In this section, we discuss two of these generalizations.

7.1 Generalization to Multiple Hidden Layers. So far, we have applied AGREL only to three-layer networks. We will now discuss how AGREL can be used if there are more than three layers. This is an important issue for any neurobiological learning scheme, because there are many levels between input and output in the cortex (Felleman & Van Essen, 1991). We change our three-layer network to a four-layer network by inserting a new input layer I before layer X (see Figure 4). Layer X thereby becomes an additional hidden layer. This modification does not change the synaptic update rules for connections w_{jk} and v_{ij}. The challenge is to define an update rule for the new layer of connections u_{hi} between layers I and X. Feedback from layer Y should guide the plasticity of connections u_{hi}. Thus, units in layer Y that provide feedforward input to layer Z should also provide feedback to layer X (see Figure 4A). However, the feedback signal is required to be different from the feedforward activation of the next layer. This can be implemented by using a separate feedback pathway (FB) where activity differs from the feedforward pathway (FF) (see Figure 4B). The separation between feedforward and feedback signals is consistent with neurobiology, because in the cortex, there are different neurons within the same cortical column that project to higher and lower levels (Felleman & Van Essen,

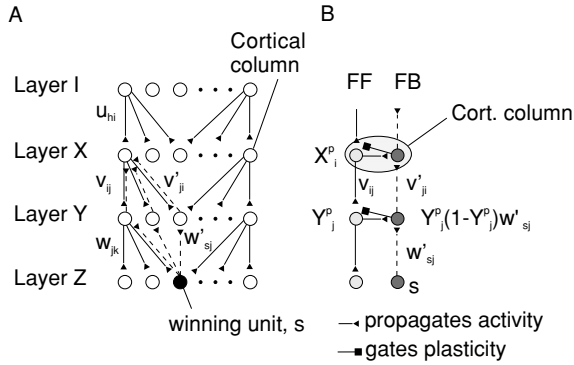

Figure 4: Generalization of AGREL to networks with more than three layers. (A) Feedforward connections u, v, and w propagate activity from the input layer I through two hidden layers to the output layer Z. The winning output unit, s, feeds back to units in layer Y through connections w'_{sj}. All units in Y that receive feedback from Z propagate it to layer X through feedback connections v'_{ji}. (B) Units of AGREL are hypothesized to correspond to cortical columns that contain FF neurons (light gray circles) that propagate activity to the next higher layer as well as FB neurons (dark gray) that propagate activity to the previous layer. FB neurons gate plasticity in the FF pathway, but they do not directly influence the activity of FF neurons (connection with square).

1991). We therefore identify a unit in AGREL with such a cortical column and assume that the activity of FF neurons differs from the activity of FB neurons.

In AGREL, feedback connections should influence only the activity of FB neurons and should not directly influence the activity of the FF neurons. All the activity in the feedback pathway originates from the winning unit s in the output layer. This unit feeds back to FB neurons in layer Y, which in turn feed back to FB neurons in layer X (see the dashed connections in Figure 4B). Once the competition in the output layer has settled, the winning unit s has activity 1 and the amount of feedback received by FB neurons in column j of layer Y equals w'_{sj} (see equation 4.6). In addition to this feedback, the FB neurons also receive an input from FF neurons of the same column, and their activity is set to $Y_j^p(1 - Y_j^p)w'_{sj}$ (see Figure 4B). The FB neurons in turn propagate $Y_j^p(1 - Y_j^p)w'_{sj}v'_{ji}$ to FB neurons in column i of layer X. All FB neurons in X feed back to column i, and the total feedback arriving in this column, $fb^p_{X_i}$, equals

$$fb^p_{X_i} = \sum_{j=1}^{M} Y_j^p \left(1 - Y_j^p\right) w'_{sj} v'_{ji}. \tag{7.1}$$

Plasticity of connections u_{hi} between layer I and X is gated by $fb^p_{X_i}$:

$$\Delta u_{hi} = \beta I^p_h X^p_i f(\delta)\big[fb^p_{X_i}(1 - X^p_i)\big], \tag{7.2}$$

the equivalent of equation 4.5. It is not difficult to show that the weight changes Δu_{hi} in AGREL are again, on average, the same as in BP. Equations 7.1 and 7.2 can be applied recursively to compute the weight changes in networks with any number of hidden layers. Thus, in the case of more than three layers, a separate feedback network is required, but this network propagates activity, not error signals, and it is therefore relatively straightforward to generalize AGREL for the training of networks with multiple layers.

7.2 AGREL as an Actor-Critic Method. On correct trials, it is essential to have a good estimate of δ, the difference between the amount of reward that was obtained and that was expected. There is little doubt that deviations from the reward expectancy are computed in the brain. If the reward is delivered immediately after the animal's response, then the reward-evoked response of dopamine neurons in the midbrain depends on reward like δ in AGREL (Fiorillo et al., 2003; Morris et al., 2004). There are various methods to compute δ on rewarded trials. In the above, we computed δ as $1 - \Pr(Z^p_{c_p} = 1)$, that is, on the basis of the a priori probability that unit c_p would win the competition, and suggested that $\Pr(Z^p_{c_p} = 1)$ can be determined by evaluating activity in the output layer at the start of the competition.

Actor-critic models (Sutton & Barto, 1998) provide an alternative method to determine reward expectancy. They are composed of two structures. The first is the Actor, which implements the mapping of sensory states onto actions. The second is the Critic, which assigns a value to every sensory state. In general, the advantage of this design is that the Critic can assign positive values to sensory states that are not associated with immediate reward but predict that reward will be obtained in the future. The reward prediction error δ is positive for any action that causes a transition to a sensory state with a higher value, and negative if the succeeding state has a lower value. Thereby, such models can also learn to choose actions that are not rewarded immediately but will be rewarded in the near future. In other words, Actor-Critic methods permit a solution to the temporal credit assignment problem. Here we investigate if AGREL can be implemented as an Actor-Critic method in the case of immediate reward. We consider the extension of AGREL to tasks with delayed rewards to be a topic for future research.

We trained an Actor-Critic model to perform the face classification task of Figure 2. AGREL is designed to map stimuli onto responses, which is the task of the Actor. We now also added a Critic network that evaluates the value of the stimulus represented by the input layer. This is illustrated

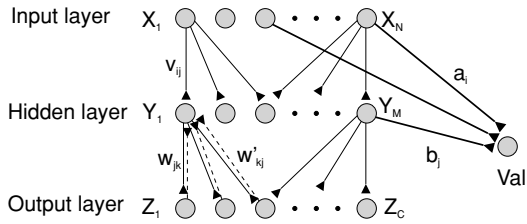

Figure 5: AGREL as an Actor-Critic method. The network includes a value estimation unit (Critic) that receives connections a_i and b_j from all units of the input layer and the hidden layer. This unit estimates the average amount of reward Val_p that is obtained for the stimulus.

in Figure 5, where we introduced a single Critic unit that estimates the expected amount of reward by linear approximation (as in Montague et al., 1995; Suri & Schultz, 2001):

$$val_p = \sum_{i=0}^{N} a_i X_i^p + \sum_{j=1}^{M} b_j Y_j^p. \tag{7.3}$$

Here, a_0 is the bias of the value-estimation unit. Now the prediction error δ is determined by a comparison between r, the amount of reward received after the trial, and the amount that was predicted for the present stimulus, val_p:

$$\delta = r - val_p. \tag{7.4}$$

δ may differ from -1 on unrewarded trials, and we therefore redefine $f(\delta)$ as follows:

$$f(\delta) = \begin{cases} \delta/(1-\delta); & \delta \geq 0 \\ -1; & \delta < 0 \end{cases}. \tag{7.5}$$

The connections a_i and b_j to the Critic unit have to be updated continuously, since modifications of synaptic weights v_{ij} and w_{jk} change the network's policy (i.e., the input-output mapping) and thereby the average amount of reward obtained for each of the patterns (Sutton & Barto, 1998). Plasticity of connections a_i and b_j is determined by

$$\Delta a_i = \alpha X_i^p \delta \tag{7.6}$$

$$\Delta b_j = \alpha Y_j^p \delta. \tag{7.7}$$

Again, all information required to update these synapses is available locally. The factor α determines the learning rate of the Critic and was set to 0.03. The learning rate β for the other synaptic weights that determine the network's output (the Actor) equaled 0.1.

With these parameters, the network required an average of 62 (S.D. = 25) presentations of the stimulus set (621 trials) to reach criterion in the face discrimination task (in 24 simulations). This is comparable to the results described above (see section 6.1). The Actor-Critic model caused an amplified representation of task-relevant features (just as in Figure 2D). The average selectivity index for the diagnostic features was 0.24, which was significantly ($p < 10^{-7}$) larger than the average selectivity index for the nondiagnostic features, which was 0.17. These results indicate that AGREL can indeed be implemented as an Actor-Critic model in the simplified case of immediate reward delivery.

8 Discussion

AGREL is a new theory for learning in classification tasks. It is the first learning scheme that is at the same time biologically plausible and as powerful as widely used but biologically implausible strategies for training artificial neural networks (see, e.g., Crick, 1989). AGREL's computational power derives from two factors known to influence synaptic plasticity (see Figure 6). The first factor is a global reward-related signal, which reaches all synapses and is presumably implemented in the brain by the release of neuromodulators. The second factor is a site-specific effect due to the feedback of neuronal activity, which assigns credit to sensory neurons that play a critical role in the selection of an action. These two factors are combined at the individual

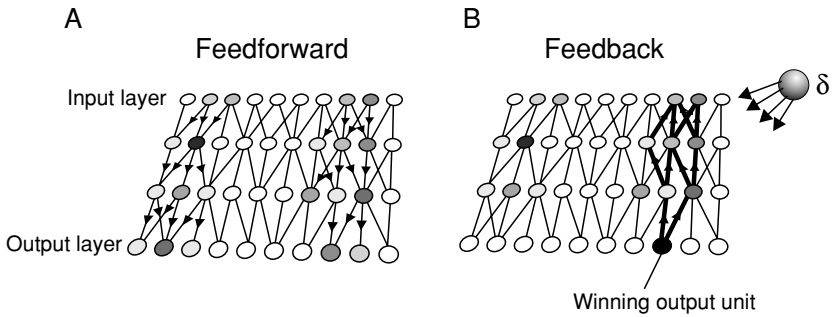

Figure 6: Two factors that modulate Hebbian plasticity. (A) An input pattern activates neurons in the various layers of the network. (B) One of the output units wins the competition. This unit feeds back to earlier processing levels. Synaptic plasticity is gated by two factors: (1) the feedback signal and (2) the global reinforcement signal δ.

synapse, where they modulate Hebbian plasticity. Here, we first discuss the neurophysiological evidence for these two factors and then how they interact at the level of the individual synapse. We then compare AGREL to other learning theories and conclude by considering current limitations of AGREL and possible directions for future research.

8.1 Global Reinforcement Signal δ. AGREL computes a signal δ that equals the difference between the amount of reward that is expected on a trial and the amount that is actually obtained. In this respect, the model follows other reinforcement learning theories (Montague et al., 1995; Suri & Schultz, 2001). There is substantial evidence that the brain computes a signal like δ. Schultz and coworkers demonstrated that the activity of dopamine neurons in the midbrain (substantia nigra and ventral tegmental area) is determined by the difference between the amount of reward expected and obtained (Ljungberg et al., 1993; Schultz et al., 1997; Schultz & Dickinson, 2000; Waelti et al., 2001). Dopamine neurons are also sensitive to increases in rewards that are predicted for the near future, and in this respect the activity of dopamine neurons fulfills the requirements of a signal that can be used in temporal difference learning (Dayan & Balleine, 2002; Montague & Berns, 2002; Schultz, 2002). Here we studied the case that rewards are delivered immediately after the response, and in that more restricted setting, the dopamine neurons have a particularly high response when the animal unexpectedly receives a reward (Fiorillo et al., 2003; Morris et al., 2004) as is required by AGREL.

The exact method used by the brain to compute δ has not yet been resolved. Neurophysiological evidence indicates that there are dedicated neuronal circuits that continuously monitor the amount of reward that is expected in the near future (Ljungberg et al., 1993; Waelti et al., 2001). Here we outlined an alternative method to compute δ that is based on the prior probability of choosing a particular category, $\Pr(Z_k^p = 1)$, which can be computed at the start of the competition between response alternatives. If the network's choice turns out to be correct, then $\Pr(Z_k^p = 1)$ equals $\Pr(Z_{c_p}^p = 1)$, the probability of a correct response, and it also equals the expected amount of reward. On incorrect trials, however, plasticity in AGREL does not depend on reward expectancy (the Actor-Critic network of section 7.2 does require this information, however; see equations 7.6 and 7.7). This prediction of AGREL could be tested in future neurophysiological experiments.

All plastic synapses have to be informed about the value of δ. In the cortex, this can be achieved by the release of neuromodulators (Pennartz, 1995; Schultz, 2002). Here we will consider two candidate neuromodulators that may carry out this job, dopamine and acetylcholine. Dopamine neurons in the substantia nigra and in the ventral tegmental area release dopamine in various nuclei of the striatum and also in regions of cortex (Garris et al., 1999). Thus, after unexpected rewards (i.e., if $\delta > 0$), the dopamine concentration is increased in these structures (Phillips, Stuber, Heien, Wightman,

& Carelli, 2003; Schultz, 2002). AGREL and other reinforcement learning models predict that this increased dopamine concentration should facilitate synaptic plasticity. Several studies support this prediction. Dopamine has been shown to be necessary for the potentiation of synapses in the prefrontal cortex, amygdala, and striatum (Gurden, Takita, & Jay, 2000; Rosenkranz & Grace, 2002; Reynolds, Hyland, & Wickens, 2001), as well as for the depression of synapses in the prefrontal cortex (Otani, Auclair, Desce, Roisin, & Crepel, 1999). Moreover, artificial stimulation of dopamine neurons in combination with an auditory stimulus expands the representation of that stimulus in auditory cortex (Bao, Chan, & Merzenich, 2001). This expansion does not occur if dopamine receptors are blocked. These results, taken together, indicate that dopamine may indeed modulate synaptic plasticity on the basis of a difference between the expected reward and the reward that was obtained.

Acetylcholine is the second candidate neuromodulator that can globally influence synaptic plasticity. It is supplied to the cortex by a number of cholinergic cell groups in the basal forebrain and brainstem (reviewed by Pennartz, 1995). Acetylcholine innervation has a relatively homogeneous density across the cortex, and it can modulate synaptic plasticity in all cortical areas. Acetylcholine differs in this respect from dopamine, as dopamine innervation is most pronounced in prefrontal cortex and amygdala and much weaker in other regions of cortex. Many studies have demonstrated that acetylcholine plays a critical role in learning and plasticity. Lesions of cholinergic input to the cortex impair learning in rats and monkeys (Winkler, Suhr, Gage, Thal, & Fisher, 1995; Easton, Ridley, Baker, & Gaffan, 2002; Warburton et al., 2003). Moreover, cholinergic lesions block synaptic plasticity in the visual cortex, as well as in the somatosensory cortex (Bear & Singer, 1986; Juliano, Ma, & Eslin, 1991). Increased cholinergic activity, on the other hand, enhances synaptic plasticity. If an auditory stimulus is paired with artificial stimulation of cholinergic nuclei, then the representation of this stimulus is expanded in auditory cortex (Bakin & Weinberger, 1996; Kilgard & Merzenich, 1998). These results indicate that acetylcholine gates the plasticity of synapses in many, if not all, areas of the cortex.

Thus, there are at least two neuromodulators, dopamine and acetylcholine, that may gate the plasticity of cortical synapses. At present, it seems too early to decide whether it is acetylcholine or dopamine, or a combination of these neuromodulators, that informs the cortical synapse about δ during learning.

8.2 Feedback from the Winning Output Unit. The global reinforcement signal by itself has little specificity to guide synaptic plasticity. AGREL therefore uses a second, site-specific factor to assign credit to those hidden units that are responsible for the selected action. This is achieved by a feedback signal from the winning output unit to the units that made it win. This feedback signal distinguishes AGREL from previous learning theories. Feedback

gates the plasticity of the lower layers, so that only the connections onto the appropriate hidden units are modified. We showed that feedback causes the hidden units to be tuned to features that are most useful for the task at hand. Indeed, our simulation results demonstrate that AGREL can account remarkably well for the changes in tuning in sensory areas that are induced by training in a categorization task. It causes a selective representation of task-relevant features (see Figure 2) and sharpens tuning at the boundary between categories (see Figure 3), just as is observed in the visual cortex of monkeys (Freedman et al., 2001; Schoups et al., 2001; Sigala & Logothetis, 2002; Baker et al., 2002).

When stimuli enter into sensory areas of the cortex, feedforward connections rapidly propagate activity to association areas and then to areas involved in response selection and execution (see Figure 6A) (Lamme & Roelfsema, 2000). The active neurons in motor cortex typically represent many different, and even incompatible, motor programs (Goldberg & Segraves, 1987; Schall & Hanes, 1993). These conflicts are resolved by a competitive interaction among the motor programs, where eventually one of them wins and suppresses the others (Seidemann, Arieli, Grinvald, & Slovin, 2002; Schall & Hanes, 1993). This competition is a stochastic process, so that trials with the same sensory stimulus can nevertheless yield different behavioral outcomes (for a computational model of action selection see, e.g., Usher & McClelland, 2001; Gold & Shadlen, 2001). We used the softmax rule to select one of the actions. This rule can be implemented in neurobiologically realistic circuits (Douglas, Koch, Mahowald, Martin, & Suarez, 1995; Nowlan & Sejnowski, 1995), and it is compatible with the computational models of action selection.

In AGREL, the winning motor program feeds its activity back to lower hierarchical levels. Anatomical studies show that feedforward and feedback connections between areas are largely reciprocal (Felleman & Van Essen, 1991; Salin & Bullier, 1995). Thus, if neurons have strong feedforward connections to neurons in a higher area, they usually also receive strong feedback from these cells. The learning rules of AGREL enforce reciprocity, as they change the strengths of feedforward and feedback connections in the same way. The consequence of reciprocity is that neurons that provide the strongest excitation to a particular motor program also receive the strongest feedback if this motor program happens to win the competition (see Figure 6B). This explains why the feedback signal can be used to assign credit to neurons in lower areas that are responsible for the choice of action. A recent neurophysiological study directly tested the specificity of feedback (Moore & Armstrong, 2003). Electrical stimulation was used to enhance the activity of neurons in the frontal eye fields (area FEF), a region of cortex involved in the generation of eye movements. During electrical stimulation, neuronal activity was recorded in area V4, which is a visual area that projects to area FEF. Electrical stimulation of FEF enhanced the activity of neurons in V4, but only if the V4 receptive

fields overlapped with the receptive fields of the stimulated FEF neurons. This is direct proof that feedback connections from neurons in a higher area project back to the neurons that are responsible for their feedforward activation.

This role of feedback is supported by many other studies. If a visual stimulus becomes the target for an eye movement, for example, neurons in the visual cortex with a receptive field at the stimulus location increase their activity. This eye movement related response enhancement is observed in areas of the parietal cortex (Colby, Duhamel, & Goldberg, 1996; Gottlieb, Kusunoki, & Goldberg, 1998), the inferotemporal cortex (Chelazzi, Miller, Duncan, & Desimone, 1993), area V4 (Boch & Fischer, 1983; Moore, 1999), and even in the primary visual cortex, that is, at the lowest hierarchical level (Supèr, van der Togt, Spekreijse, & Lamme, 2004). Psychophysical data indicate that this enhancement of neuronal responses is a correlate of visual attention (Desimone & Duncan, 1995). It is indeed well established that attention is directed to items that become the target of an eye movement. Coupling between attention and eye movements is so strong that observers are virtually unable to visually discriminate a target at one location just before they execute an eye movement to another location (Hoffman & Subramaniam 1995; Kowler, Anderson, Dosher, & Blaser, 1995; Deubel & Schneider, 1996). The strong coupling between the intention to move and the shift of attention to the features that instruct the movement is also known as the "premotor theory of attention" (Rizzolatti, Riggio, & Sheliga, 1994). We conjecture that this coupling is explained by the reciprocity of feedforward and feedback connections. The influence of response selection on the distribution of visual attention is what psychologists call goal-driven or top-down attention. Here we proposed a new role of this top-down attentional signal, which is to enable plasticity at earlier processing levels (see the thick connections in Figure 6B). We showed how this feedback signal amplifies the representation of diagnostic features in sensory areas. AGREL thereby increases the saliency of relevant features in perception, and this would correspond to what psychologists call an effect on stimulus-driven (bottom-up) attention. Thus, the theory predicts a new interaction between top-down and stimulus-driven attention: by its influence on plasticity, goal-driven attention eventually increases the saliency of diagnostic features in the course of training.

Support for the effect of goal-driven attention on plasticity was obtained in a psychophysical study by Ahissar and Hochstein (1993). They presented the same visual stimuli to two groups of human observers. One group was asked to report about one attribute of the stimuli, and the other group had to report about a different attribute. The subjects' ability to observe differences in these attributes improved during training. However, this improvement occurred only for the attribute that they were asked to report about; performance for the attribute that was not attentively practiced remained constant. Similar results were obtained in monkeys that were trained in the orientation discrimination task of Figure 3. Discrimination performance at

the trained location became much better than performance at the untrained location. After the training, V1 neurons with receptive fields at the trained location had a sharper orientation tuning, even though the bottom-up input at the two locations had been similar during training. These results, taken together, demonstrate that attentional feedback indeed gates the plasticity of sensory representations.

In this study, we modeled only the effect of feedback on synaptic plasticity. Weight changes in AGREL will deviate from those of BP if the effect of feedback on neuronal activity is included in the model. Neurophysiological studies in visual cortical areas demonstrated that attended objects typically evoke responses that are 20% to 40% stronger than those evoked by nonattended objects (Moran & Desimone, 1985; Chelazzi et al., 1993; Motter, 1993; Schall & Hanes, 1993; Treue & Maunsell, 1996; Luck, Chelazzi, Hillyard, & Desimone, 1997; Roelfsema, Lamme, & Spekreijse, 1998; Reynolds, Pasternak, & Desimone, 2000). In an additional simulation of the face categorization task (see Figure 2), we investigated whether this influences the convergence of AGREL. The strength of the effect of feedback on activity was set such that excitatory feedback connections increased the responses of units in the hidden layer by an average of 19% (maximum 57%), and inhibitory feedback connections decreased the responses by an average of 17% (maximum 98%). The network learned the task after an average of 653 trials (24 simulations), which is comparable to the convergence in the absence of the effect of feedback on activity (613 trials; U-test, $p > 0.2$). Thus, in this simulation, the effect of feedback on activity did not deteriorate learning. Moreover, neurophysiological studies indicate that there is a substantial fraction (20–50%) of visual neurons that is not influenced by attention and that therefore always carries the unperturbed feedforward response (just like the FF pathway of Figure 4B) (e.g., Moran & Desimone, 1985; Motter, 1993; Treue & Maunsell, 1996; Luck et al., 1997; Roelfsema et al., 1988; Roelfsema, Lamme & Spekreijse, 2004; Reynolds et al., 2000). Thus, the pure sensory response is always available in the various areas of the visual cortex.

Szabo, Almeida, Deco, and Stetter (in press) proposed that an effect of feedback on neuronal activity could explain the enhanced representation of diagnostic features over nondiagnostic features in area IT. Future neurophysiological studies should be able to distinguish between effects of training on feedback connections (as proposed by Szabo et al., in press) and the effects of training on feedforward connections to area IT (as in AGREL). On the one hand, if changes in the feedforward connections are responsible for the enhanced tuning to diagnostic features, then this effect should be visible at the start of the visual response of IT neurons. On the other hand, if the enhanced representation of diagnostic features is due to altered feedback, then it should not occur during the initial visual response but rather after an additional delay imposed by the loop through the response selection stage (Sugase, Yamane, Ueno, & Kawano, 1999; see also Lamme & Roelfsema, 2000).

8.3 Interactions Between Feedforward Connections, Neuromodulators, and Feedback. AGREL proposes a specific set of interactions between pre- and postsynaptic activity, feedback effects, and neuromodulators at the level of the individual synapse. Many of the hypothesized interactions have not yet been addressed experimentally. However, there are a few exceptions. In AGREL, plasticity depends on a multiplicative interaction between feedforward input and feedback (see equations 4.6 and 7.2). A number of studies provide support for such a multiplicative interaction, although they measured activity, not synaptic plasticity. The first is the electrical stimulation experiment in area FEF discussed above (Moore & Armstrong, 2003). In this experiment, electrical stimulation in FEF only influenced V4 neurons with a visual stimulus in their receptive field, and not neurons with an empty receptive field. Thus, feedback by itself does not activate neurons; rather, it amplifies activity provided by the sensory input. This also holds true for attentional effects. If attention is drawn to a stimulus, it enhances the response of neurons that are activated by this stimulus. Attention has no effect on neurons that are not driven by the stimulus (McAdams & Maunsell, 1999; Treue & Martínez Trujillo, 1999). These results demonstrate that neuronal activity depends on a multiplicative interaction between the feedforward and feedback input.

Receptor pharmacology suggests a possible mechanism for the effect of feedback on neuronal activity (Dehaene, Sergent, & Changeux, 2003). A neuron's initial feedforward response is mainly driven by its α-amino-3-hydroxy-5-methyl-4-isoxazole propionate (AMPA) receptors, whereas cortical feedback strongly activates N-methyl-D-aspartate (NMDA) receptors (Salin & Bullier, 1995; Shima & Tanji, 1998). The opening of NMDA channels is voltage dependent so that current flows through these channels only if the cell is sufficiently depolarized by its AMPA receptors (Collingridge & Bliss, 1987). This explains why NMDA hardly influences spontaneous activity if applied to the visual cortex at low concentrations (Fox, Sato, & Daw, 1990). Apparently the neurons are insufficiently depolarized to open NMDA channels if there is no stimulus in their receptive field. If a visual stimulus drives the neurons, however, the same concentration of NMDA increases activity by an amount proportional to the strength of the visual response. In other words, NMDA increases the neuron's sensory gain, and its blockers reduce the gain (Fox et al., 1990). Thus, NMDA-receptor activation has a multiplicative effect on the strength of the visual response.

The pharmacology of glutamate receptors can, at the same time, also explain how feedback connections gate the plasticity of feedforward connections (see equations 4.5 and 7.2). The first step in many forms of synaptic plasticity is the entry of calcium in the postsynaptic neurons through NMDA channels. This calcium activates various biochemical cascades that upregulate the efficacy and number of AMPA receptors (Muller, Joly, & Lynch, 1988; Shi et al., 1999). This raises the exciting possibility that NMDA-ergic feedback connections might gate the plasticity of AMPA-ergic feedforward connections.

8.4 Comparison to Other Learning Theories. We first compare AGREL to BP and then to other biologically inspired learning schemes. The discovery of BP in the 1970s and early 1980s was a major breakthrough in the field of neural networks. BP was the first learning scheme that could efficiently train neural networks with hidden layers (Werbos, 1974; Rumelhart et al., 1986). Hidden layers are important, since they greatly expand the potential of neural networks. If the task is classification, for example, networks without a hidden layer can separate input patterns only along linear category boundaries (Bishop, 1995). Networks with one or more hidden layers, on the other hand, can find solutions for a much larger class of categorization problems, because they can form nonlinear category boundaries (the XOR problem is a well-known case where this is necessary). BP ensures that the output of hidden units becomes tuned to the appropriate nonlinear combinations of input unit activations, and it can thereby solve these additional categorization problems. However, BP is implausible from a neurobiological perspective (Crick, 1989). Its first implausible feature is that it requires a separate pathway for the backpropagation of error signals. The required error signals are different at each site of plasticity (Zipser & Rumelhart, 1990; Crick, 1989). These site-specific error signals are not observed in the cortex. Instead, feedback connections cause site-specific attentional effects that can be observed in many, if not all, cortical areas (reviewed by Lamme & Roelfsema, 2000).

The second implausible feature of BP is that it uses a teacher. It is unlikely that a teacher exists in the brain that can supply the target activity of all neurons in the motor cortex after each trial. Moreover, animals can learn without a teacher by exploring the various response alternatives. If they make an error, they simply try out another response on the next trial. Reinforcement learning theories, such as AGREL, are designed to learn by trial and error. A drawback of learning without a teacher is that it takes more time, and the training of a neural network with AGREL is slower than with BP, especially if there are many response alternatives or if the probability of choosing the correct response is small. We conjecture, however, that this must be true for any learning scheme that has to find the correct response by trial and error.

There are a number of previous theories that were inspired by the implausibility of BP and that suggested learning rules closer to biology. These previous theories can be broadly divided into two classes: (1) theories that compute a local error signal at each individual hidden unit and (2) other reinforcement learning theories, which use a global error signal that is broadcast to all sites of plasticity. In our discussion, we will not consider previous studies that combined reinforcement learning with error backpropagation (e.g., Tesauro, 1995) as they were not concerned with biological plausibility.

A number of studies suggested methods to compute the local error signal at each hidden unit as is required by BP. A straightforward way is to use a dedicated error network with neurons that propagate the error signal from the output layer back to lower network levels (Zipser and Rumelhart,

1990). Körding & König (2001) suggested an alternative, and somewhat speculative, possibility that a single neuron might propagate feedforward activation to higher levels as well as an error signal to lower levels by using different types of action potentials. In their proposal, the bottom-up input to a neuron drives normal action potentials, whereas the top-down error signal activates calcium spikes. However, there is no evidence that calcium spikes are tuned to the BP error. Another crucial limitation of the theories of Zipser and Rumelhart (1990) and Körding & König (2001) is that they did not get rid of the teacher of BP.

Another method to compute the error gradient at each hidden unit was suggested by O'Reilly (1996) in his generalized recirculation algorithm (*GeneRec*). Like AGREL, *GeneRec* recirculates activity between layers, which are interconnected with feedforward and feedback connections. Moreover, feedforward connections and feedback connections in *GeneRec* have a similar strength, which aids in assigning credit to hidden units that are responsible for the stimulus-response mapping, just as in AGREL. There are, however, also important differences between the two learning schemes. In *GeneRec*, learning takes place by the alternation of two phases, a "minus" and a "plus" phase. In the minus phase, output units are activated by other units in the network. In the plus phase, the teacher determines the output units' activity. *GeneRec* changes synaptic weights considering four factors: the pre- and postsynaptic activity in the minus phase and the pre- and postsynaptic activity in the plus phase. This implies that all units must remember their activity of the minus phase while the network is running in the plus phase. It is unclear how this could be implemented in the cortex. Another important drawback of *GeneRec* is that it requires a teacher to specify the target activity of all output units.

Reinforcement learning represents the second class of theories that provides learning rules that might be implemented in the brain (Pennartz, 1995; Sutton & Barto, 1998; Suri & Schultz, 2001). The advantage of these theories is that learning can take place while the system explores the various response alternatives by trial and error. Many of the previous reinforcement learning theories provide solutions to the temporal credit assignment problem that arises when rewards are delivered after a delay, or when the animal first has to progress though a sequence of states before reward delivery (Montague et al., 1995; Sutton & Barto, 1998; Baxter & Bartlett, 2001). One way to solve the temporal credit assignment problem is to use a separate critic network that assigns a hedonic value to sensory states. The critic network can be trained with temporal difference learning to assign high hedonic values to states predicting that reward will be obtained in the near future (Sutton & Barto, 1998). Although we have not yet tested AGREL in tasks with delayed rewards, we showed that it is compatible with such a separate critic network. We stress, however, that AGREL was designed to improve the *actor* of an Actor-Critic model and that we leave the question of whether feedback connections can improve learning in the critic network for future research.

AGREL borrows important concepts from previous reinforcement learning theories; in particular, it adopts the reward prediction error δ. A number of earlier studies demonstrated that δ can be used to guide synaptic plasticity in the network that maps sensory inputs onto actions (actor-network). In some of these studies, δ was used to train networks with two layers (Barto & Anandan, 1985; Montague et al., 1995). If δ is the only signal that influences synaptic plasticity, the definition of a learning rule for networks with three or more layers is more problematic, because the global signal has insufficient specificity to resolve the spatial credit assignment problem. Nevertheless, it is possible to train multilayer feedforward networks just on the basis of δ and pre- and postsynaptic activity. One algorithm that has been used is the associative reward-penalty algorithm (A_{R-P}) (Barto, 1985; Barto & Anandan, 1985; Mazzoni, Andersen, & Jordan, 1991), and another class is formed by REINFORCE algorithms (Williams, 1992). Hidden units in these algorithms are stochastic and randomly choose between an active and inactive state. They attempt to estimate their contribution to the output by correlating their own behavior with the reward, without knowledge about their impact on the output layer (see also Seung, 2003). Thus, connections onto hidden units that are not involved in a decision may also change after a trial. AGREL differs in this respect, since it modifies connections onto only hidden units that were involved in the decision. We compared AGREL to published data obtained with A_{R-P} in a small network with a single hidden unit (Barto, 1985) and found that convergence in AGREL is three times faster. We predict that this factor increases if there are many hidden units between the input and output layer so that it becomes more important to assign the credit to the correct ones.

There is a further difference between AGREL and REINFORCE. In AGREL, the average change in synaptic strength is proportional to the gradient on the cross-entropy $-\nabla Q_p$ (Q_p is defined in equation 3.3). In contrast, REINFORCE makes weight changes that are proportional to $\nabla E^p(r)$, where $E^p(r)$ is the average amount of reward obtained with stimulus p (Williams, 1992). For the tasks considered here, there is a simple relationship between the two gradients, since

$$\begin{aligned}\frac{\partial E^p(r)}{\partial w_{jc_p}} &= \frac{\partial P\left(Z^p_{c_p}=1\right)}{\partial w_{jc_p}} = Y^p_j P\left(Z^p_{c_p}=1\right)\left\{1-P\left(Z^p_{c_p}=1\right)\right\}\\ &= -P\left(Z^p_{c_p}=1\right)\frac{\partial Q_p}{\partial w_{jc_p}}, \quad \text{and} \qquad (8.1)\\ \frac{\partial E^p(r)}{\partial w_{jk}} &= \frac{\partial P\left(Z^p_{c_p}=1\right)}{\partial w_{jk}} = -Y^p_j P\left(Z^p_{c_p}=1\right)P\left(Z^p_k=1\right)\\ &= -P\left(Z^p_{c_p}=1\right)\frac{\partial Q_p}{\partial w_{jk}}; \quad k\neq c_p, \qquad (8.2)\end{aligned}$$

and similarly

$$\frac{\partial E^p(r)}{\partial v_{ij}} = \frac{\partial P\left(Z^p_{c_p} = 1\right)}{\partial v_{ij}} = -P\left(Z^p_{c_p} = 1\right)\frac{\partial Q_p}{\partial v_{ij}}. \tag{8.3}$$

Thus, in general,

$$\nabla E^p(r) = -P\left(Z^p_{c_p} = 1\right) \cdot \nabla Q_p. \tag{8.4}$$

The two gradients therefore have the same direction for each pattern p, but REINFORCE makes on average small steps in weight space for patterns that are usually classified erroneously (i.e., small $P(Z^p_{c_p} = 1)$) and larger steps for stimuli that are often classified correctly. The total gradient equals the sum across the gradients for the individual patterns and differs between AGREL and REINFORCE:

$$-\nabla Q = \sum_p -\nabla Q_p, \tag{8.5}$$

whereas

$$\nabla E(r) = \sum_p \nabla E^p(r) = \sum_p -P\left(Z^p_{c_p} = 1\right) \cdot \nabla Q_p. \tag{8.6}$$

This difference between REINFORCE and AGREL was readily apparent when we compared the two algorithms on benchmark problems. REINFORCE often failed to leave regions of weight space where one or more input patterns had a small $P(Z^p_{c_p} = 1)$, whereas AGREL usually succeeded in training the network.

8.5 Limitations and Future Extensions. So far, we have applied AGREL only to tasks where the animal learns to associate a unique action with every input pattern and where the reward is delivered immediately after a correct response. Moreover, we always rewarded one of a limited number of potential actions; all other actions were not rewarded. Future work will have to determine whether AGREL can be extended to more complex situations that have been addressed by other reinforcement learning theories. First, it will be important to investigate whether AGREL is compatible with tasks where rewards are delivered after a delay and sequential decision tasks where the animal progresses through a number of states before reward delivery (Sutton & Barto, 1998; Dayan & Balleine, 2002). We made a first step in this direction by implementing AGREL as an Actor-Critic method. However, we have yet to study the behavior of AGREL in tasks with delayed reward delivery. Also, we did not investigate tasks where rewards are

delivered probabilistically or where payoffs are variable. Second, it will be important to extend AGREL to regression tasks, where the output is a continuous variable rather than the choice of one of a limited set of categories. Animals are commonly confronted with tasks that require regression, for example, if they have to learn to reach to locations in space.

9 Conclusion

We conclude that the inclusion of an attentional feedback signal in reinforcement learning permits new, biologically plausible learning rules that are as efficient as error BP in forming useful internal representations. Gating of plasticity by a combination of reinforcement signals and attentional feedback permits learning rules where the average changes in synaptic weights are precisely in the direction of the BP error gradient. AGREL thereby establishes a new link between supervised learning and biologically inspired reinforcement learning theories, theories of learning that were largely unconnected in the past.

Acknowledgments

We thank Carl van Vreeswijk and Cyriel Pennartz for helpful comments on an earlier version of the manuscript. P.R.R. was supported by a HFSP Young Investigators grant.

References

Ahissar, M., & Hochstein, S. (1993). Attentional control of early perceptual learning. *Proc. Natl. Acad. Sci. USA*, *90*, 5718–5722.

Baker, C. I., Behrmann, M., & Olson, C. R. (2002). Impact of learning on representation of parts and wholes in monkey inferotemporal cortex. *Nat. Neurosci.*, *5*, 1210–1216.

Bakin, J. S., & Weinberger, N. M. (1996). Induction of a physiological memory in the cerebral cortex by stimulation of the nucleus basalis. *Proc. Natl. Acad. Sci. USA*, *93*, 11219–11224.

Bao, S., Chan, V. T., & Merzenich, M. M. (2001). Cortical remodelling induced by activity of ventral tegmental dopamine neurons. *Nature*, *412*, 79–81.

Barto, A. G. (1985). Learning by statistical cooperation of self-interested neuron-like computing elements. *Human Neurobiol.*, *4*, 229–256.

Barto, A. G., & Anandan, P. (1985). Pattern-recognizing stochastic learning automata. *IEEE Trans. Syst., Man. and Cybernet.*, *SMC-15*, 360–375.

Baxter, J., & Bartlett, P. L. (2001). Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, *15*, 319–350.

Bear, M. F., & Singer, W. (1986). Modulation of visual cortical plasticity by acetylcholine and noradrenaline. *Nature*, *320*, 172–176

Becker, S., & Hinton, G. E. (1992). Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, *355*, 161–163.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford: Oxford University Press.

Boch, R., & Fischer, B. (1983). Saccadic reaction times and activation of the prelunate cortex: Parallel observations in trained rhesus monkeys. *Exp. Brain Res.*, *50*, 201–210.

Chelazzi, L., Miller, E. K., Duncan, J., & Desimone, R. (1993). A neural basis for visual search in inferior temporal cortex. *Nature*, *363*, 345–347.

Colby, C. L., Duhamel, J. R., & Goldberg, M. E. (1996). Visual, presaccadic, and cognitive activation of single neurons in monkey lateral intraparietal area. *J. Neurophysiol.*, *76*, 2841–2852.

Collingridge, G. L., & Bliss, P. (1987). NMDA receptors—their role in long-term potentiation. *Trends Neurosci.*, *10*, 288–293.

Crick, F. (1989). The recent excitement about neural networks. *Nature*, *337*, 129–132.

Dayan, P., & Balleine, B. W. (2002). Reward, motivation, and reinforcement learning. *Neuron*, *36*, 285–298.

Dehaene, S., Sergent, C., & Changeux, J. P. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proc. Natl. Acad. Sci. USA*, *100*, 8520–8525.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.*, *18*, 193–222.

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Res.*, *36*, 1827–1837.

Douglas, R. J., Koch, C., Mahowald, M., Martin, K. A. C., & Suarez, H. H. (1995). Recurrent excitation in neocortical circuits. *Science*, *269*, 981–985.

Easton, A., Ridley, R. M., Baker, H. F., & Gaffan, D. (2002). Unilateral lesions of the cholinergic basal forebrain and fornix in one hemisphere and inferior temporal cortex in the opposite hemisphere produce severe learning impairements in rhesus monkeys. *Cereb. Cortex*, *12*, 729–736.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex*, *1*, 1–47.

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, *299*, 1898–1902.

Fox, K., Sato, H., & Daw, N. (1990). The effect of varying stimulus intensity on NMDA-receptor activity in cat visual cortex. *J. Neurophysiol.*, *64*, 1413–1428.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, *291*, 312–316.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2003). A comparison of primate prefrontal cortex and inferior temporal cortices during visual categorization. *J. Neurosci.*, *15*, 5235–5246.

Garris, P. A., Kilpatrick, M., Bunin, M. A., Michale, D., Walker, Q. D., & Wightman, R. M. (1999). Dissociation of dopamine release in the nucleus accumbens from intracranial self-stimulation. *Nature*, *398*, 67–69.

Gold, J. I., & Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends Cogn. Sci.*, *5*, 10–16.

Goldberg, M. E., & Segraves, M. A. (1987). Visuospatial and motor attention in the monkey. *Neuropsychologia*, *25*, 107–118.

Gorman, R. P., & Sejnowski, T. (1988). Analysis of hidden units in a layered network trained to classify sonar targets. *Neural Networks*, *1*, 75–89.

Gottlieb, J. P., Kusunoki, M., & Goldberg, M. E. (1998). The representation of visual salience in monkey parietal cortex. *Nature*, *391*, 481–484.

Gurden, H., Takita, M., & Jay, T. M. (2000). Essential role of D1 but not D2 receptors in the NMDA receptor-dependent long-term potentiation at hippocampal-prefrontal cortex synapses in vivo. *J. Neurosci., 20*, RC106 (1–5).

Gustafsson, B., & Wigstrom, H. (1988). Physiological mechanisms underlying long-term potentiation. *Trends Neurosci.*, *11*, 156–162.

Hinton, G. E., Dayan, P., Frey, B. J., & Neal, R. M. (1995). The "wake-sleep" algorithm for unsupervised neural networks. *Science, 268*, 1158–1161.

Hoffman, J. E., & Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Percept. Psychophys.*, *57*, 787–795.

Juliano, S. L., Ma, W., & Eslin, D. (1991). Cholinergic depletion prevents expansion of topographic maps in somatosensory cortex. *Proc. Natl. Acad. Sci. USA*, *88*, 780–784.

Kilgard, M. P., & Merzenich, M. M. (1998). Cortical map reorganization enabled by nucleus basalis activity. *Science, 279*, 1714–1718.

Körding, K. P., & König, P. (2001). Supervised and unsupervised learning with two sites of synaptic integration. *J. Comp. Neurosci.*, *11*, 207–215.

Kowler, E., Anderson, E., Dosher, B., & Blaser, E. (1995). The role of attention in the programming of saccades. *Vision Res.*, *35*, 1897–1916.

Lamme, V. A. F., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.*, *23*, 571–579.

Ljungberg, T., Apicella, P., & Schultz, W. (1993). Responses of monkey dopamine neurons during learning of behavioral reactions. *J. Neurophysiol.*, *67*, 145–163.

Luck, S. J., Chelazzi, L., Hillyard, S. A., & Desimone, R. (1997). Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *J. Neurophysiol.*, *77*, 24–42.

Malinow, R., & Miller, J. P. (1986). Postsynaptic hyperpolarisation during conditioning reversibly blocks induction of long-term potentiation. *Nature*, *320*, 529–530.

Mazzoni, P., Andersen, R. A., & Jordan, M. I. (1991). A more biologically plausible learning rule for neural networks. *Proc. Natl. Acad. Sci. USA*, *88*, 4433–4437.

McAdams, C. J., & Maunsell, J. H. R. (1999). Effects of attention on orientation-tuning functions of single neurons in macaque area V4. *J. Neurosci.*, *19*, 431–441.

Montague, P. R., & Berns, G. S. (2002). Neural economics and the biological substrates of valuation. *Neuron*, *36*, 265–284.

Montague, P. R., Dayan, P., Person, C., & Sejnowski, T. J. (1995). Bee foraging in uncertain environments using predictive Hebbian learning. *Nature, 377*, 725–728.

Moore, T. (1999). Shape representations and visual guidance of saccadic eye movements. *Science, 285*, 1914–1917.

Moore, T., & Armstrong, K. M. (2003). Selective gating of visual signals by microstimulation of frontal cortex. *Nature*, *421*, 370–373.

Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science*, *229*, 782–784.

Morris, G., Arkadir, D., Nevet, A., Vaadia, E., & Bergman, H. (2004). Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. *Neuron*, *43*, 133–143.

Motter, B. C. (1993). Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli. *J. Neurophysiol., 70*, 909–919.

Muller, D., Joly, M., & Lynch, G. (1988). Contributions of quisqualate and NMDA receptors to the induction and expression of LTP. *Science, 242*, 1694–1697.

Nowlan, S. J., & Sejnowski, T. J. (1995). A selection model for motion processing in area MT of primates. *J. Neurosci., 15*, 1195–1214.

O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Comp., 8*, 895–938.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*, 607–609.

Otani, S., Auclair, N., Desce, J. M., Roisin, M. P., & Crepel, F. (1999). Dopamine receptors and groups I and II mGluRs cooperate for long-term depression induction in rat prefrontal cortex through converging postsynaptic activation of MAP kinases. *J. Neurosci.*, 15, 9788–9802.

Pennartz, C. A. M. (1995). The ascending neuromodulatory systems in learning by reinforcement: Comparing computational conjectures with experimental findings. *Brain Res. Rev., 21*, 219–245.

Phillips, P. E. M., Stuber, G. D., Heien, M. L. H. V., Wightman, R. M., & Carelli, R. M. (2003). Subsecond dopamine release promotes cocaine seeking. *Nature, 422*, 614–618.

Reynolds, J. H., Pasternak, T., & Desimone, R. (2000). Attention increases sensitivity of V4 neurons. *Neuron, 26*, 703–714.

Reynolds, J. N. J., Hyland, B. I., & Wickens, J. R. (2001). A cellular mechanism of reward-related learning. *Nature, 413*, 67–70.

Rizzolatti, G., Riggio, L., & Sheliga, B. M. (1994). Space and selective attention. In C. Umiltà & M. Moscovitch (Eds.), *Attention and performance XV. Conscious and nonconscious information processing* (pp. 231–265). Cambridge, MA: MIT Press.

Roelfsema, P. R., Lamme, V. A. F., & Spekreijse, H. (1998). Object-based attention in the primary visual cortex of the macaque monkey. *Nature, 395*, 376–381.

Roelfsema, P. R., Lamme, V. A. F., & Spekreijse, H. (2004). Synchrony and covariation of firing rates in the primary visual cortex during contour grouping. *Nature Neurosci., 7*, 982–991.

Rosenkranz, J. A., & Grace, A. A. (2002). Dopamine-mediated modulation of odour-evoked amygdala potentials during Pavlovian conditioning. *Nature, 417*, 282–287.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group (Eds.), *Parallel distributed processing* (Vol. 1, pp. 318–364). Cambridge, MA: MIT Press.

Rumelhart, D. E., & Zipser, D. (1986). Feature discovery by competitive learning. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group (Eds.), *Parallel distributed processing* (Vol. 1, pp. 151–193). Cambridge, MA: MIT Press.

Salin, P. A., & Bullier, J. (1995). Corticocortical connections in the visual system: Structure and function. *Physiol. Rev., 75*, 107–154.

Schall, J. D., & Hanes, D. P. (1993). Neural basis of saccade target selection in frontal eye field during visual search. *Nature, 366*, 467–469.

Schoups, A., Vogels, R., Qian, N., & Orban, G. A. (2001). Practising orientation identification improves orientation coding in V1 neurons. *Nature, 412*, 549–553.

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron*, *36*, 241–263.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275*, 1593–1599.

Schultz, W., & Dickinson, A. (2000). Neuronal coding of prediction errors. *Annu. Rev. Neurosci., 23*, 473–500.

Seidemann, E., Arieli, A., Grinvald, A., & Slovin, H. (2002). Dynamics of depolarization and hyperpolarization in the frontal cortex and saccade goal. *Science, 295*, 862–865.

Seung, H. S. (2003). Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. *Neuron*, *40*, 1063–1073.

Shi, S.-H., Hayashi, Y., Petralia, R. S., Zaman, S. H., Wenthold, R. J., Svoboda, K., & Malinow, R. (1999). Rapid spine delivery and redistribution of AMPA receptors after synaptic NMDA receptor activation. *Science*, *284*, 1811–1816.

Shima K., & Tanji J. (1998). Involvement of NMDA and non-NMDA receptors in the neuronal responses of the primary motor cortex to input from the supplementary motor area and somatosensory cortex: Studies of task-performing monkeys. *Jpn. J. Physiol., 48*, 275–290.

Sigala, N., & Logothetis, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature, 415*, 318–320.

Sugase, Y., Yamane, S., Ueno, S., & Kawano, K. (1999). Global and fine information coded by single neurons in the temporal visual cortex. *Nature*, *400*, 869–873.

Supèr, H., van der Togt, C., Spekreijse, H., & Lamme, V. A. F. (2004). Correspondence of presaccadic activity in the monkey primary visual cortex with saccadic eye movements. *Proc. Natl. Acad. Sci. USA*, *101*, 3230–3235.

Suri, R. E., & Schultz, W. (2001). Temporal difference model reproduces anticipatory neural activity. *Neural Comp.*, *13*, 841–862.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Szabo M., Almeida R., Deco G., & Stetter, M. (in press). A neuronal model for the shaping of feature selectivity in IT by visual categorization. *Neurocomputing*.

Tanaka, K. (1995). Neuronal mechanisms of object recognition. *Science*, *262*, 685–688.

Tesauro, G. (1995). Temporal difference learning and TD-Gammon. *Comm. ACM*, *38*, 58–68.

Treue, S., & Martínez Trujillo, J. C. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, *399*, 575–579.

Treue, S., & Maunsell, J. H. R. (1996). Attentional modulation of visual motion processing in cortical areas MT and MST. *Nature*, *382*, 539–541.

Usher, M., & McClelland, J. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychol. Rev.*, *108*, 550–592.

van Ooyen, A., & Nienhuis, B. (1992). Improving the convergence of the backpropagation algorithm. *Neural Networks*, *5*, 465–471.

Vogels, R., & Orban, G. A. (1990). How well do response changes in striate neurons signal differences in orientation? A study in the discriminating monkey. *J. Neurosci.*, *10*, 3543–3558.

Waelti, P., Dickinson, A., & Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature, 412*, 43–48.

Warburton, E. C., Koder, T., Cho, K., Massey, P. V., Duguid, G., Barker, G. R. I., Aggleton, J. P., Bashir, Z. I., & Brown, M. W. (2003). Cholinergic neurotransmission is essential for perirhinal cortical plasticity and recognition memory. *Neuron*, *38*, 987–996.

Werbos, P. J. (1974). *Beyond regression: New tools for prediction and analysis in the behavioral sciences*. Unpublished doctoral dissertation, Harvard University.

Winkler, J., Suhr, S. T., Gage, F. H., Thal, L. J., & Fisher, L. J. (1995). Essential role of neocortical acetylcholine in spatial memory. *Nature*, *375*, 484–487.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, *8*, 229–256.

Zipser, D., & Rumelhart, D. E. (1990). The neurobiological significance of the new learning models. In E. L. Schwartz (Ed.), *Computational neuroscience* (pp. 192–200). Cambridge, MA: MIT Press.

Received April 22, 2004; accepted March 22, 2005.