HIPPOCAMPUS 13:494–524 (2003)

Comparison of Computational Models of Familiarity Discrimination in the Perirhinal Cortex

Rafal Bogacz[1] and Malcolm W. Brown[2*]

[1]*Department of Computer Science, University of Bristol, Bristol, United Kingdom*
[2]*MRC Centre for Synaptic Plasticity, Department of Anatomy, University of Bristol, Bristol, United Kingdom*

ABSTRACT: This study compares the efficiency and plausibility of published computational models of familiarity discrimination in the perirhinal cortex. Substantial evidence indicates that the perirhinal cortex is involved in both the familiarity discrimination aspect of recognition memory and in perceptual functions involved with representations of complete stimuli (i.e., object identification). Published models of how the perirhinal cortex may perform familiarity discrimination can be divided into two groups. The first group assumes that a proportion of perirhinal neurons form a network specialised just for familiarity discrimination (these models may be based on Hebbian or anti-Hebbian synaptic plasticity). In contrast, the second group assumes that both familiarity discrimination and learning representations of complete stimuli are performed within a single combined network. This study establishes that when the responses of neurons that provide input to the familiarity discrimination network are correlated (as indicated by experimental data), specialised networks based on anti-Hebbian learning may recognise the previous occurrence of many more stimuli (i.e., have a capacity up to thousands of times larger) than specialised networks based on Hebbian learning. The currently published combined models do not learn an optimal stimulus representation (they do not fully extract statistically independent features), and hence their capacities are even lower than those of the specialised models based on Hebbian learning. Hence, the combined models published thus far are critically less efficient than the specialised models based on anti-Hebbian learning. This study also compares the consistency of the models with experimental observations concerning what is known of synaptic plasticity in the perirhinal cortex and the responses of its neurons. Many theoretically important parameters remain undetermined, and experiments are suggested to provide information critical for refining and distinguishing between the various models. However, the above theoretical arguments and currently published data favour the existence of a separate network specialised for familiarity discrimination. *Hippocampus 2003;13: 494–524.* © 2003 Wiley-Liss, Inc.

KEY WORDS: recognition memory; neural network models; novelty detection; feature extraction; hippocampal region

INTRODUCTION

Work in monkeys has established that discrimination of the relative familiarity or novelty of visual stimuli is dependent on the perirhinal cortex. This finding is consistent with studies of amnesic patients (Eichenbaum et al., 1994; Aggleton and Shaw, 1996; Murray, 1996; Suzuki, 1996; Brown and Xiang, 1998; Buffalo et al., 1998; Aggleton and Brown, 1999; Murray and Bussey, 1999; Brown and Aggleton, 2001). Thus, damage to the perirhinal cortex results in impairments in recognition memory tasks that rely on discrimination of the relative familiarity of objects (Murray, 1996; Brown and Aggleton, 2001). Moreover, within the monkey's perirhinal cortex, ~25% of neurons respond strongly to the sight of novel objects but respond only weakly or briefly when these objects are seen again (Brown et al., 1987; Riches et al., 1991; Fahy et al., 1993; Li et al., 1993; Miller et al., 1993; Sobotka and Ringo, 1993; Brown and Xiang, 1998; Xiang and Brown, 1998). Analysis of the population of such responses attests to very fast discrimination of the novelty or familiarity of stimuli: response differences occur within 100 ms of stimulus onset (Miller et al., 1993; Xiang and Brown, 1998). This finding accords with the ability of human subjects to make such discriminations rapidly (Seeck et al., 1997; Hintzman et al., 1998). In addition, the population of these neuronal responses manifests a very large storage capacity, as the responses of individual neurons continue to signal the novelty or familiarity of objects even when many hundreds of objects have been seen (Fahy et al., 1993; Li et al., 1993; Xiang and Brown, 1998). This finding is in accordance with the huge capacity of human recognition memory. Standing (1973) examined this capacity using single trial learning and forced-choice recognition. After seeing 10,000 pictures subjects could recognise them with an average accuracy of 83%. Furthermore, the number of stimuli retained in the recognition memory as a function of the amount of material presented followed a power law, which led Standing (1973, p 207) to con-

Grant sponsor: Overseas Research Scholarship; Grant sponsor: Wellcome Trust; Grant sponsor: Biotechnology and Biological Science Research Council; Grant sponsor: Medical Research Council.
Rafal Bogacz is currently at the Department of Applied and Computational Mathematics, Princeton University, Princeton, NJ.
*Correspondence to: M.W. Brown, Department of Anatomy, University of Bristol, Bristol BS8 1TD, UK. E-mail: m.w.brown@bristol.ac.uk
Accepted 20 June 2002
DOI 10.1002/hipo.10093

clude: "The capacity of recognition memory for pictures is almost limitless."

The perirhinal cortex has other functions, including involvement in the perception and categorisation of complex stimuli (Murray and Bussey, 1999). Thus, damage to the perirhinal cortex results in impairment in tasks requiring perception of whole stimuli (e.g., distinguishing between complex objects) but does not lead to impairment in tasks requiring perception of only individual features (e.g., distinguishing between objects of different colours) (Buckley and Gaffan, 1998). Therefore, it has been suggested that perirhinal cortex neurons represent conjunctions of features of visual stimuli, perhaps resulting in representation of complete stimuli ("objects"), whereas regions earlier in the visual processing stream contain neurons that represent simple features from which these complex conjunctions are formed (Murray and Bussey, 1999). Accordingly, there is evidence that the perirhinal cortex is involved in both the familiarity discrimination aspect of recognition memory and perceptual functions involved in representations of complete stimuli. This study discusses whether perirhinal neurons belong to a single network that performs combined feature extraction and familiarity discrimination or to two separable networks, one specialised for familiarity discrimination and another for perceptual feature extraction.

The published models of how the perirhinal cortex may perform familiarity discrimination can be divided in two groups. Models of the first group assume that a proportion of the perirhinal neurons form a network specialised just for familiarity discrimination, and are referred to as the specialised models (Bogacz et al., 2001a,b). In contrast, models of the second group assume that both familiarity discrimination and learning representations of complete stimuli are performed within a single network. They are referred to as combined models (Sohal and Hasselmo, 2000; Norman and O'Reilly, 2001).

Bogacz et al. (1999, 2001a) showed that, under specific conditions, specialised networks can achieve very high storage capacity. If the perirhinal cortex worked similarly to the specialised models, it alone could discriminate the familiarity of many more stimuli than current neural network models indicate could be recalled (recollected) by all the remaining areas of the cerebral cortex. This efficiency and speed of detecting novelty provide an evolutionary advantage, providing a reason for the existence of a familiarity discrimination network in addition to networks used for recollection. The networks combining familiarity discrimination with learning representations have been simulated, but it remains to be established whether they, too, can achieve such efficiency.

This study compares the efficiency and consistency with experimental observations of the proposed specialised and combined models of familiarity discrimination in the perirhinal cortex. The capacity achieved by various models is calculated and compared with the results of simulations. The analysis of the biological plausibility of the models and their efficiency is then used to indicate types of neuronal networks that might perform familiarity discrimination in the perirhinal cortex.

The present report is presented in three major sections, followed by Conclusions and an Appendix section. The first major section, Description of Networks That Can Perform Familiarity Discrimination, introduces and reviews all the published models of familiarity discrimination in the perirhinal cortex. The second major section, Comparison of the Models' Efficiencies, compares the capacity of the models. The third major section, Consistency of the Models With Experimental Observations, compares the consistency of the models with experimental observations and suggests further experiments. The derivations of capacity and details of simulations are given in the Appendices. Accordingly, the main text of this study is presented with minimal use of equations so as to be understandable without a computational modelling background. The first and second major sections after the Introduction focus on modelling computations performed by novelty neurons, the ~10% of perirhinal neurons that respond strongly to the first presentations of novel stimuli, but only briefly or weakly to presentations of previously seen stimuli (Xiang and Brown, 1998; Brown and Xiang, 1998).

DESCRIPTION OF NETWORKS THAT CAN PERFORM FAMILIARITY DISCRIMINATION

As introduced above, the models of familiarity discrimination in the perirhinal cortex may be divided in two groups: specialised models assuming that a proportion of perirhinal neurons form a network specialised just in familiarity discrimination (Bogacz et al., 2001a,b), and combined models assuming that familiarity discrimination and learning representations of complete stimuli are performed within a single network (Sohal and Hasselmo, 2000; Norman and O'Reilly, 2001). Hence, the division between the specialised and the combined models is based on their function as given by their authors, rather than on the other features of the models (e.g., their architecture or assumed plasticity). The two groups of models will now be briefly described; additional information, including the mathematical description of the simulated versions of the models is presented in Appendix A.

For ease of explanation and mathematical analysis, the networks are introduced using a simple model of neurons, similar to that of McCulloch and Pitts (1943). This model does not consider changes of the membrane potentials of neurons in time. The original models proposed by Norman and O'Reilly (2001) and Sohal and Hasselmo (2000) were simulated with continuous neurons. Therefore, the present study analyses simplified versions of these models, although the simplifications do not invalidate the major conclusions. We assume that each visual stimulus is represented by a specific pattern of activity of the neurons providing input to the familiarity discrimination network and that the activities of these input neurons represent features of the stimuli. We further assume that the neurons providing input to the network may be in one of two states: active or inactive. For example, after presentation of a visual stimulus, the active state of an input neuron corresponds to an increase in its activity (i.e., a response) and the inactive state to no increase (i.e., to no response). However, as demonstrated previously (Bogacz et al., 2001a), it is possible to extend a model of a perirhinal network based on binary neurons to a model based on

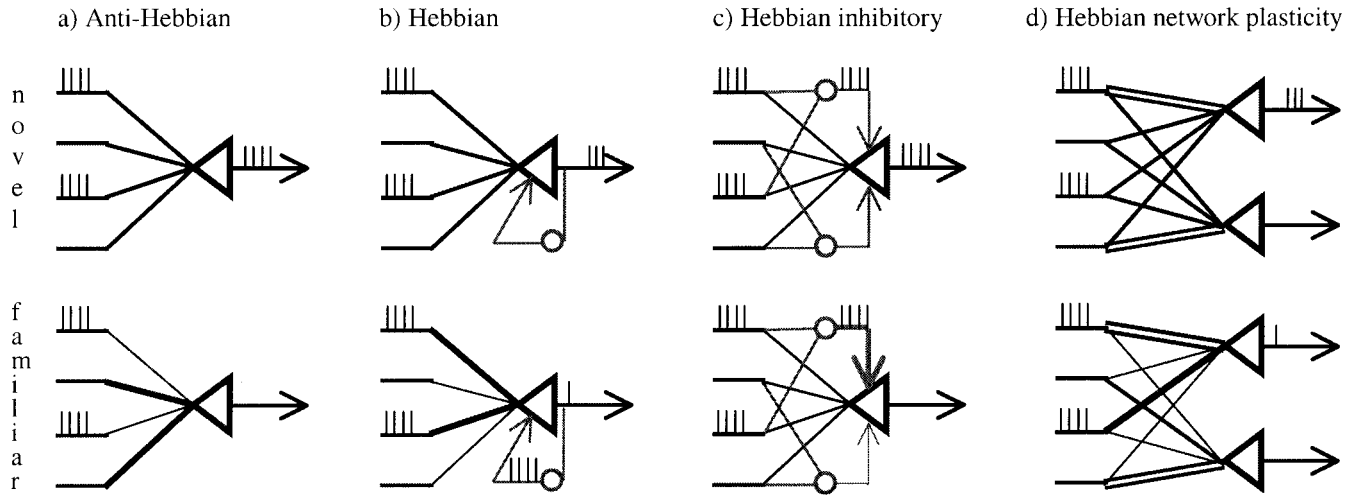

FIGURE 1. Synaptic and network mechanisms that may underlie the decrease of perirhinal neurons' response for familiar stimuli. In each panel, the triangle represents an excitatory novelty neuron (Xiang and Brown, 1998; Brown and Xiang, 1998), and the circle represents an inhibitory interneuron. Lines on the left side of each panel denote inputs to the network, which are axons of neurons whose activity encodes visual stimuli. "Spikes" over the lines indicate that the corresponding neuron is active, a lack of spikes, that it is inactive. The thickness of the lines indicates the strength of the synaptic connections. The top row of panels illustrates synaptic weights and neuronal responses for a novel stimulus, and the lower row of panels when this stimulus is presented again (i.e., for a familiar stimulus). Three mechanisms are illustrated based on (a) anti-Hebbian learning, (b) Hebbian learning, and (c) Hebbian learning between inhibitory interneurons and novelty neurons. d: Synaptic weight modification in the Hebbian model. For simplicity, the inhibitory neurons (mentioned in the text and shown in b) are not shown. After presentation of a novel stimulus, the number of active novelty neurons is limited (only the upper one is active), for example by nonmodifiable connections with high synaptic weights (denoted by double lines, for simplicity only one is shown for each neuron). The synaptic weights of the active novelty neurons are modified as in panel b, while the weights of the inactive neurons are modified in the opposite way, e.g., the synaptic weight from the active input to the inactive novelty neuron is decreased as if by homosynaptic LTD.

more realistic spiking neurons (Gerstner, 1998), with the operational principles, capacity, and efficiency remaining essentially unchanged.

Specialised Models

Three specialised models are described, each differing in the assumed synaptic plasticity of the synapses on to the novelty neurons. First, the plasticity of a single neuron from each model is described, followed by a description of how the neurons may be combined into a network.

Synaptic plasticity of novelty neurons

A proportion of neurons in the perirhinal cortex have weaker responses after presentation of familiar stimuli than of novel stimuli (Brown et al., 1987; Riches et al., 1991; Fahy et al., 1993; Li et al., 1993; Miller et al., 1993; Sobotka and Ringo, 1993; Brown and Xiang, 1998; Xiang and Brown 1998). A number of synaptic and network mechanisms may underlie this decrease of response; three of these mechanisms are illustrated in Figure 1a–c (see the section, Independent Responses of Novelty Neurons, for a discussion of Fig. 1d).

Figure 1a presents the anti-Hebbian model (Bogacz and Brown, 2002; Brown and Xiang 1998; Kohonen, 1989) i.e., one based on decreases rather than increases in synaptic strength. After presentation of a novel stimulus, the synaptic weights of connections from active input neurons are decreased as if by homosynaptic long-term depression (LTD) (Kemp and Bashir, 2001). This synaptic modification decreases the sum of the synaptic weights of the novelty neuron. Hence to maintain the overall excitability of the neuron, the synaptic weights of connections from inactive input neurons must be increased (Fig. 1a). When the same stimulus is presented again, the membrane potential of the novelty neuron will be lower (because the weights of synapses of inputs that were active for this stimulus have been reduced), and the novelty neuron will be inactive (or, more generally, less active). Thus, the neuron responds more strongly to novel stimuli than to familiar stimuli.

Figure 1b shows the Hebbian model (Bogacz et al., 1999, 2001a) based on Hebbian synaptic plasticity. After presentation of a novel stimulus, synaptic weights from active inputs are increased as if by long-term potentiation (LTP) (Bliss and Collingridge, 1993), while weights from inactive units are decreased as if by heterosynaptic LTD (Ito, 1989). These changes produce an initially higher response of novelty neurons for familiar stimuli than for novel. However, in the network, the novelty neurons project to inhibitory neurons; the result is a higher level of inhibition for familiar than for novel stimuli, and the increased inhibition leads to a smaller neuronal response for familiar stimuli than for novel stimuli (Fig. 1b).

Figure 1c shows the Hebbian inhibitory model based on Hebbian learning in synapses connecting inhibitory interneurons to novelty neurons. It assumes that presentation of a visual stimulus produces a unique pattern of activity across inhibitory neurons, as well as in the excitatory inputs. After presentation of a novel stim-

ulus, synaptic weights from active inhibitory neurons to active novelty neurons are increased, while weights from inactive inhibitory neurons are decreased. Therefore, for subsequent presentations of the same stimulus, the novelty neuron will receive more inhibition and hence decrease its response.

One could also design a model in which synaptic strengths (weights) from inputs to the inhibitory interneurons are modified according to Hebbian rules. Thus, inhibition would be increased as a result of stronger input to the inhibitory neuron for familiar than for novel stimuli, rather than as a result of strengthening the output synapses of the inhibitory neuron. It is also possible to create models combining two of the above models, in particular a model combining the anti-Hebbian and Hebbian inhibitory models, in which the occurrences of familiar stimuli are stored in both modifiable excitatory (from input neurons) and inhibitory (from inhibitory neurons) synapses of the novelty neurons. However, for simplicity, such complex models are not analysed in this study.

Independent responses of novelty neurons

Each of the familiarity discrimination models includes a single layer of novelty neurons receiving projections from the input neurons. If each novelty neuron makes its own decision about stimulus familiarity, the overall response ("answer") of the network is encoded in the population activity of the novelty neurons. In each of the models, it is necessary to ensure that individual novelty neurons remain independent assessors of familiarity if the information storage capacity of the network is to be maximised (Bogacz et al., 2001a). Otherwise, should all the novelty neurons be active after the presentation of each of a series of novel stimuli, then the synaptic weights of each of the novelty neurons would be modified in the same way, and hence all the novelty neurons would come to have highly correlated weights. Thus, eventually, they would all be active or inactive together and the whole network would have the same capacity as a single novelty neuron. To avoid this problem, the number of novelty neurons active for any one stimulus must be limited; that is, only a subset of novelty neurons must respond to any given stimulus.

There are at least two means of limiting the number of active novelty neurons. The first means is inhibitory competition: only the fraction of neurons with the highest membrane potentials are selected to be active, the activity of the remainder being suppressed by inhibition, and only these most active neurons have their weights modified (Norman and O'Reilly, 2001). This method of limiting the number of active novelty neurons is used in the combined models (based on Hebbian learning) (see the section, Combined Models) and in the anti-Hebbian model.

The second method of ensuring this selectivity of response of the novelty neurons is to provide specific connections with high synaptic weights from the network inputs to subsets of novelty neurons. Although this method requires the additional assumption of the existence of specialised connections (and therefore may seem less plausible), it makes mathematical analysis of network behaviour simpler. Therefore, the Hebbian model as analysed in this study assumes that the number of active neurons is limited by specific connections with high synaptic weights. As the more plausible models that limit the number of active novelty neurons by competition are much more difficult to analyse, for these models only approximate expressions for capacity may be found mathematically. However, many properties that may be proved mathematically for the Hebbian model with strong connections are also valid for other familiarity discrimination networks based on Hebbian learning (Bogacz and Brown, 2002).

In fact, experiments show that novelty neurons are stimulus selective, as required by the models (Xiang and Brown, 1998; Li et al., 1993; Miller et al., 1993). However, note that for a network specialised for familiarity discrimination (and not learning representations), this selectivity is required solely to increase the efficiency of the network, and not because the implied representation of the visual stimuli is used for some further processing within the network (such representations provide the assumed inputs to the familiarity discrimination network). Nevertheless, the activity of the different groups of novelty neurons could potentially provide information about which of a set of perceived object(s) is novel in the case when an animal views a scene consisting of a number of objects. It should be noted that the theoretical argument concerning maximising capacity provides one explanation for the observed selectivity of perirhinal neuronal responses.

When such a network rather than a single neuron is considered in the Hebbian model described by Bogacz et al. (2001a), another synaptic change is introduced: the weights of connections between active inputs and inactive novelty neurons are reduced as if, for example, by homosynaptic LTD (Fig. 1d). This decrease is required for the decision of the network to be given by its total activity, rather than a more complex function, namely the difference between activity of the most active and the least active neurons (Bogacz and Brown, 2002), and hence makes more plausible its implementation in the real brain. The anti-Hebbian and Hebbian inhibitory models require analogous weight modifications if the decision of the network is to be given by its total activity or, for their biologically plausible implementation, may require the decision of the network to be given by a more complex function than the total activity of its neurons (Bogacz and Brown, 2002), but these details will not be further considered here. The mathematical description of the simulated versions of the Hebbian and anti-Hebbian models are contained in Appendices A.1 and A.2.

The description of the Hebbian inhibitory model (see the section, Synaptic Plasticity of Novelty Neurons) leaves undetermined the function of the excitatory connections to the novelty neurons, and how their weights might be modified. Therefore, the Hebbian inhibitory model is not simulated in this study and its capacity is not analysed. Note, however, that if we assume in the Hebbian inhibitory model that the excitatory input weights of the novelty neurons are not modified and the number of active novelty neurons is limited by competition, then the operation of the Hebbian inhibitory model is equivalent to the operation of the anti-Hebbian model, because the Hebbian learning at inhibitory synapses is equivalent to the anti-Hebbian learning at excitatory synapses. Therefore, the Hebbian-inhibitory model may potentially also achieve the same efficiency with respect to the number of synapses used for recognition memory, as does the anti-Hebbian model. However, that there are fewer inhibitory than excitatory synapses

(Thompson et al., 2001) (as is pointed out in the section, Synaptic Plasticity).

Combined Models

To understand the operation of the combined models, it is necessary first to illustrate principles underlying networks that learn stimulus representations. This is done using the simple example illustrated in Figure 2.

Learning representations

The responses of neurons in areas higher in the hierarchy of visual system processing represent more and more complicated features (Felleman and Van Essen, 1991), for example, from simple features such as changes of luminance on receptors in the retina (Meister and Berry, 1999), to the complex features of entire stimuli in perirhinal neurons (Murray and Bussey, 1999). Connections between neurons resulting in these representations are partly encoded genetically, but they are also tuned by experience (Wiesel and Hubel, 1965). Barlow (1989) suggested that the goal of sensory processing is to reduce redundancy in sensory information and to achieve a representation in which the activities of sensory neurons encode independent features. The process of learning such a representation is often referred to as feature extraction, as illustrated in Figure 2.

Many network models for feature extraction have been proposed (e.g., von der Malsburg, 1973; Grossberg, 1976; Foldiak, 1990; Olshausen and Field, 1996; Harpur and Prager, 1996; Bell and Sejnowski, 1997; Bogacz et al., 2001c). Many of these network models include Hebbian competitive learning. Such models work in the following way. After presentation of a stimulus, only a proportion of neurons, those with the highest membrane potentials are active—these are the neurons representing features that best describe that stimulus. The weights of the active neurons are then modified according to Hebbian rules; this in turn further tunes the features represented by the neurons. In this way, the weight modification is similar to that in the Hebbian model of familiarity discrimination.

Combining learning representations with familiarity discrimination

Combined models assume that the inputs to the network performing familiarity discrimination come from different areas, which encode different aspects of a stimulus. The activities of the inputs may be correlated, as by hypothesis these networks have not yet completed feature extraction.

Li et al. (1993) suggested that the reduction of the number of perirhinal neurons that are active after presentation of familiar compared with novel stimuli was caused by learning a sparse representation of the stimuli. After presentation of a novel stimulus, synaptic weights are modified such that neurons that do not represent features of the stimulus very well will not be active during subsequent presentations of the stimulus (as in the feature extraction process described above). Thus, during the process, a more precise and sparse representation of a familiar stimulus is formed

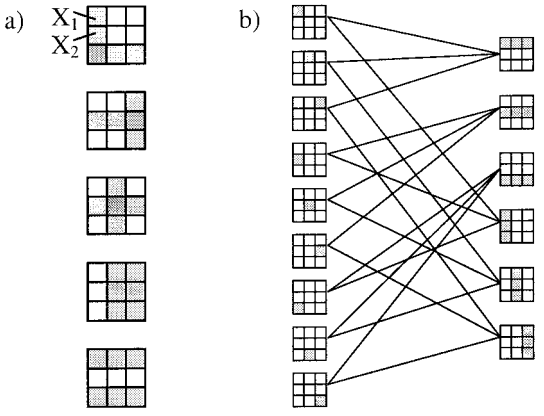

FIGURE 2. Example of correlated patterns and a network extracting features. a: Patterns consisting of randomly occurring vertical or horizontal lines of three pixels in length. If we assume that there are nine sensory neurons, each representing one pixel, for these line stimuli many pairs of neurons would have correlated, i.e., nonindependent, activity, as shown below. These sensory neurons do not achieve optimal feature extraction. The correlation between pixels X_1 and X_2 is $\frac{1}{2}$, which intuitively comes from the fact that if X_1 is on, then with probability at least $\frac{1}{2}$ there is a vertical line, so that X_2 should be on as well. This correlation may be computed formally as follows. Let us denote the active state by 1, the inactive by 0, and the probability of being active by a. Therefore, X_1 and X_2 have mean a and variance $a-a^2$. The probability of X_1 and X_2 being active simultaneously is $a(\frac{1}{2}+\frac{1}{2}a)$. Hence the covariance between X_1 and X_2 is equal to $\frac{1}{2}(a-a^2)$, so the correlation between them is $\frac{1}{2}$. b: A sample network transforming patterns from a) into independent features. This feature extraction may be achieved by appropriately connecting these neurons to a second layer of neurons. Each box of nine squares denotes a neuron, and the pattern on the box denotes the feature of a stimulus to which the neuron responds. Lines denote connections between neurons. The activities of neurons in the output (right) layer of this network encode independent features (i.e., lines), so the correlation between each pair of neurons is zero, and satisfactory feature extraction has been accomplished.

(Li et al., 1993). This idea is implemented in the models of Norman and O'Reilly (2001) and Sohal and Hasselmo (2000).

Norman and O'Reilly (2001) proposed a computational model of human recognition memory. The model includes two parts: a neocortical part responsible for the familiarity discrimination aspect of recognition memory and a hippocampal part responsible for the recollective aspect. The model proposed by Norman and O'Reilly (2001) provides a clear explanation for the results of many psychological experiments (see the section, Psychological Features). Here we analyse a simplified version of the neocortical part of the Norman and O'Reilly (2001) model, referring to it as the combined competitive model. (For a discussion of the relationship between the original Norman and O'Reilly model and the combined competitive model, see the section, Psychological Features.)

The combined competitive model is similar to the Hebbian model (Fig. 1b), with the exception of two features. First, the limitation of the number of active novelty neurons is achieved not by special strong connections, but by inhibition and competition: the active novelty neurons are those which have the highest membrane potentials. Second, only the weights of active novelty neurons are updated (i.e., there is no homosynaptic LTD, as illustrated

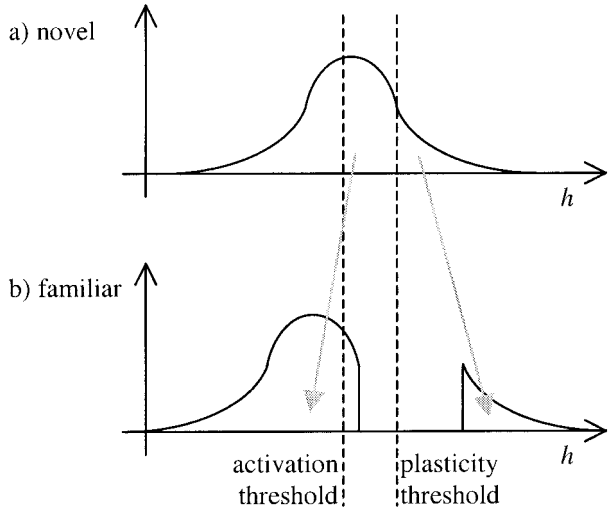

FIGURE 3. Intuitive explanation of the double threshold model. Distribution of membrane potentials for (a) a novel stimulus, and (b) a familiar stimulus. The horizontal axis denotes membrane potential and the vertical axis the number of neurons with a given membrane potential. Dashed lines show two thresholds: the activation threshold, above which neurons are active, and the plasticity threshold separating neurons whose synaptic weights are modified in different ways.

in Fig. 1d). The mathematical description of the simulated version of the combined competitive model is contained in Appendix A.3.

Sohal and Hasselmo (2000) proposed a model to explain the responses of perirhinal neurons during recognition memory tasks. Two separate mechanisms were proposed for long-term and short-term recognition memory. As this study is concerned with long-term recognition memory, only the part of the Sohal and Hasselmo (2000) model concerned with long-term familiarity discrimination is analysed here. It is termed the double threshold model.

The double threshold model also employs Hebbian rules of learning, but the decrease in the number of neurons active for familiar stimuli is not caused by inhibition. The way in which this network discriminates familiarity is illustrated in Figure 3. After presentation of a novel stimulus, the membrane potentials of the novelty neurons may be assumed to follow a normal distribution (Fig. 3a). The proportion of neurons with membrane potentials that are higher than a certain value, denoted as the plasticity threshold in Figure 3, have their synaptic weights modified as for active novelty neurons in the specialised Hebbian model (Fig. 1b): the weights from active inputs are increased and the weights from inactive inputs are decreased. The weights of neurons with membrane potentials below the plasticity threshold are modified as for inactive novelty neurons in the specialised Hebbian model (see the inactive neuron in Fig. 1d): the weights from active inputs are decreased. These weight modifications mean that when the novel stimulus is presented subsequently the membrane potentials are even higher for the neurons which were above the plasticity threshold on the first presentation (they follow the change indicated by the right arrow connecting Fig. 3a with 3b), while the membrane potentials for other neurons are even lower (they follow the left arrow in Fig. 3). In the double threshold model, there is an activation threshold—neurons with membrane potentials above this threshold are active—which is smaller than the plasticity threshold (Fig. 3). If the activation threshold is set appropriately (for example by inhibition within the network), more neurons are active for novel than for familiar stimuli (compare the areas under the distribution density curves to the right of the activation threshold in Fig. 3a,b). The number of active novelty neurons can thus be used as the familiarity criterion. The mathematical description of the double threshold model is contained in the Appendix A.4.

COMPARISON OF THE MODELS' EFFICIENCIES

This section compares the capacity for familiarity discrimination of the models. First, the capacity is presented with the simplifying assumption that activities of all the novelty neurons and the network inputs are uncorrelated (see Capacity for Uncorrelated Input, below). For this case, it is shown that all the models achieve very high capacity. Then, it is shown that the currently published combined models are unable to extract fully independent features (see Feature Extraction by Combined Models). Next, what is known about the correlation between perirhinal neurons is reviewed (see Correlation Between Responses of Real Perirhinal Neurons). It is then established that when the responses of neurons providing input to the familiarity discrimination network are correlated, the Hebbian model has a much lower capacity than the anti-Hebbian model (see Capacity for Correlated Input Patterns). The theoretical upper limit of capacity of the combined models is slightly above the capacity of the Hebbian model, but the currently published combined models do not extract statistically independent features and hence have capacities below even that for the Hebbian model. The recognition capacity of human perirhinal cortex is estimated (see Estimation of Capacity of Large Networks), and the abilities of the networks to detect unusual stimuli are analysed (see Detecting Unusual Stimuli).

Capacity for Uncorrelated Input Patterns

Storage capacity is defined as the number of presented stimuli for which a network can discriminate familiarity with an accuracy of 99%. Bogacz and Brown (2002) show that for the simplifying assumption that activities of all the novelty neurons and the network inputs are uncorrelated, all the models achieve very high storage capacity. For example, the Hebbian model has a storage capacity that is equal to 0.023 times the total number of modifiable synapses of all the novelty neurons (Bogacz and Brown, 2002). The capacity calculated by Bogacz et al. (2001a) is one-half that given here because of differences in the assumptions made concerning the precise form of the decision function/learning rule; these assumptions are discussed by Bogacz and Brown (2002). A similar capacity for uncorrelated input patterns may also be achieved by the anti-Hebbian model (Bogacz and Brown, 2002).

Both the combined competitive and the double threshold networks may also be used for familiarity discrimination in the case where inputs already encode independent features, as for the spe-

cialised models. Bogacz and Brown (2002) have shown that for this case both combined models achieve a capacity similar to that of the specialised models.

It is especially noteworthy that the capacity of the familiarity discrimination models is much greater than the capacity of associative memories for recall. The former is proportional to the number of synapses in the network, the latter to that number divided by the number of neurons (Amit, 1989). This difference may be intuitively explained by comparing the two tasks: recall—for example, you see a person and you want to recall his/her name and the episode of the previous meeting with the person—and familiarity discrimination—you see a person and you want to determine whether you have seen this person before. In the first case, the network has to recall the whole representation of the name and the episode, which is encoded in the activity of a number of neurons—let us denote this number by N. By contrast, for familiarity discrimination, there is just a binary output: the stimulus is novel or familiar. The number of outputs in the case of familiarity discrimination is N times smaller (so, in this sense, familiarity discrimination is N times easier than recall). Therefore, intuitively, the capacity for familiarity discrimination is of order N times higher.

Using estimates of the size of the human perirhinal cortex (areas 35 and 36) (Insausti et al., 1998) and assuming "idealised" noise-free neurons with uncorrelated activities, Bogacz et al. (2001a) estimated that according to the specialised familiarity discrimination models the human perirhinal cortex should be able to discriminate familiarity (with probability of error 10^{-6}) for $\sim 10^8$ stimuli. This would mean that a person living for 100 years ($\sim 3 \times 10^9$ s) who was presented with a picture every 30 s could still recognise almost all these pictures as familiar. For a 1% probability of error, the capacity of human perirhinal cortex for the above assumption would be of order of 10^9 stimuli, i.e., equivalent to storing the occurrence of a new picture every 3 s.

To summarise, this section establishes that under the assumption of uncorrelated activity of the inputs, all models can achieve very high capacity.

Feature Extraction by Combined Models

This section investigates the quality of features extracted by the published networks combining familiarity discrimination and feature extraction. This quality determines the efficiency of the combined models and, in particular, their capacity.

Although both combined competitive and double threshold models are able to extract features (Sohal and Hasselmo, 2000; Norman and O'Reilly, 2001) neither of them extracts fully independent features. Consider, as a simple example, the case of the stimuli consisting of lines, illustrated in Figure 2a. After learning (i.e., repeated presentation of stimuli consisting of lines, as those in Fig. 2a), the activity of a combined network's neurons would represent the presence of lines in the stimuli but, in the combined models as published, there is no mechanism to ensure that different neurons represent different lines. Thus, in these models as presented, feature extraction will in general be inefficient: although the neurons will learn to represent features of stimuli, one feature may be represented by a group of neurons and other features may not be represented at all. As, for instance, the responses of neurons representing the same feature are necessarily correlated, overall, the activities of the network's neurons will not be uncorrelated. Previous work has established that networks including only feedforward connections between input and output layers cannot learn to extract fully independent features. To ensure that neurons represent completely independent features, network models have to employ additional mechanisms—such as plastic inhibitory connections between neurons (e.g., Foldiak, 1990) or backprojections from feature neurons to input neurons (e.g., Olshausen and Field, 1996; Harpur and Prager, 1996).

The fact that the combined competitive and the double threshold networks do not have mechanisms to ensure that all the independent features present in the patterns are represented by novelty neurons would be likely to be a major disadvantage, since a feature carrying behaviourally relevant information might be missed. However, the probability of such a feature being missed can be reduced by increasing redundancy, i.e., by having more novelty neurons than there are independent features in the input patterns. Appendix B.1 shows that a feature extraction network such as the combined competitive and the double threshold networks, that has k times more neurons than features to be extracted, will on average miss a proportion of features given by approximately e^{-k} (where $e \approx 2.71$). For example, if there is the same number of neurons as features, such a network will miss $\sim$37% of the features (by contrast to specialised feature extraction networks that can extract a complete set of features, see Bogacz et al., 2001c). In order to miss $\leq$1% of the features, there must be $\sim$4.6 times more neurons than features.

Furthermore, in the case analysed in Appendix B.1 all the features were assumed to occur equally often and were represented by an equal number of input neurons. The behaviours of the networks change if (1) a group of features is more frequently present than the rest, (2) a group of features is represented by larger numbers of inputs than the rest, or (3) the neurons representing the features of the group are more active. In this case, in the combined competitive or the double threshold networks the features evoking greater than average activity will come to be represented by the great majority of the novelty neurons. This occurs because the stronger features "attract" the novelty neurons more strongly during feature extraction (O'Reilly and Munakata, 2000), and there is no mechanism of communication between the novelty neurons in these models to counteract this process. Hence the stronger features will be overrepresented and the weaker underrepresented.

To verify the above prediction, we analysed features extracted by a combined competitive network of 50 neurons and a double threshold network of 50 neurons. The input patterns were generated by superposition of some of 50 independent features. Each feature is represented by the activity of 5 input neurons, and each pattern is created by superposition of 5 features chosen randomly from the set of 50 features (e.g., Fig. 4a). Furthermore, 10 of these 50 independent features were stronger; that is, the neurons representing these features evoked an activity that was *strength* times larger (e.g., Fig. 4b; for details of the pattern generation, see Appendix B.2). Hence a value for *strength* = 1 corresponds to all features being equal, and larger values of *strength* correspond to the

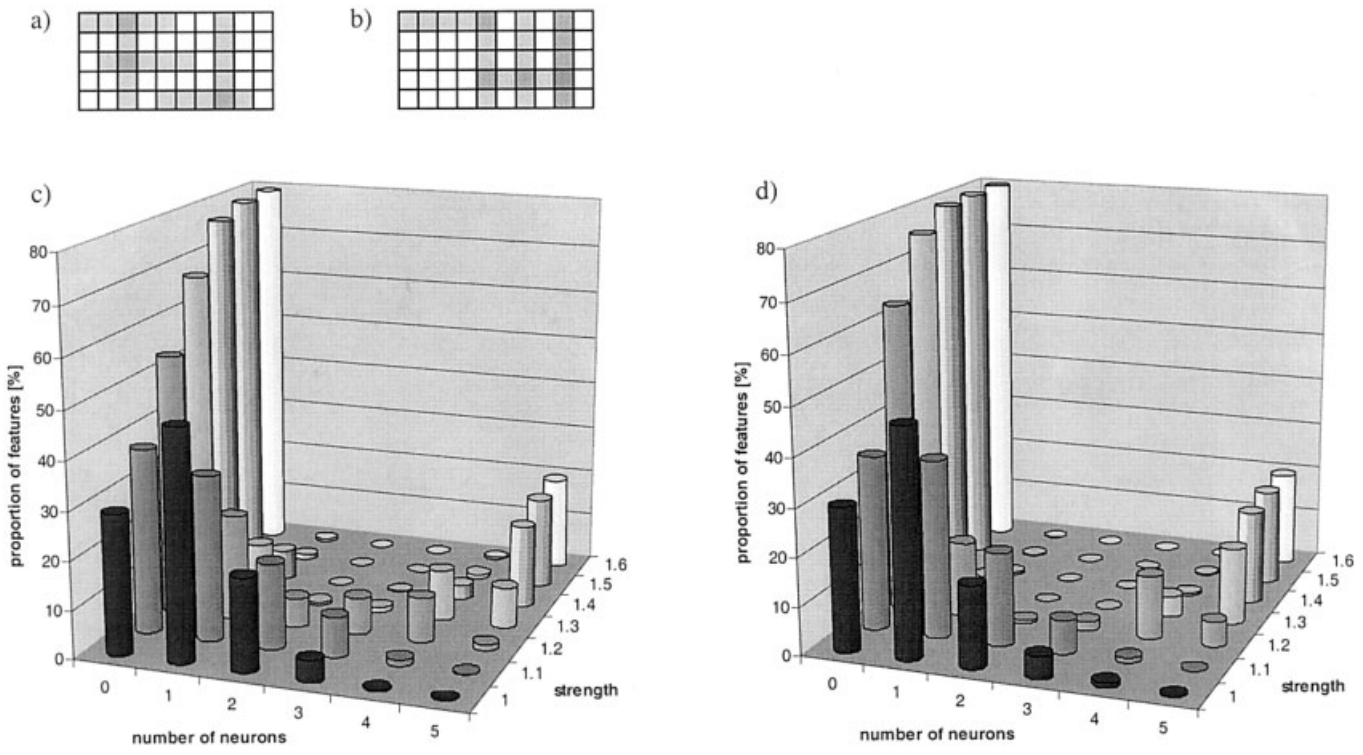

FIGURE 4. Example of feature extraction by the combined models. a,b: Sample patterns used in simulations. Darkness of each square corresponds to the level of activity of one input neuron. a: Pattern generated by superposition of equally strong features. b: One feature was stronger (the right vertical line). The features in a and b are shown as lines of 5 pixels for simplicity of explanation, but the features used in simulations were not necessarily lines, i.e., they were randomly chosen sets of five inputs (see Appendix B.2 for details) c,d: Proportions of features represented by different numbers of neurons in the combined competitive model (c) and the double threshold model (d). Simulations were performed for networks of N = 50 neurons. In each simulation session, the weights were initialised to random values then 5,000 patterns were presented and the weights modified according to the learning rules of the given model (Appendices A.3, A.4). Then the numbers of features represented by different numbers of neurons were found. For each value of the parameter *strength* for strong features (shown on y-axis), the simulations were repeated 50 times. The average proportions of features represented by different numbers of neurons (number of neurons is shown on x-axis) are plotted as bars.

10 features being stronger. After presentation of 5,000 such patterns, the neurons' weights were matched with the independent features used to generate the patterns, to find which feature was represented by which novelty neuron (see Appendix B.2). Then the numbers of features not represented by any neuron, represented by one neuron, by two, etc. were counted. The results of simulations are shown in Figure 4c,d.

Figure 4c,d shows that both combined models extract features in a similar way: for equally strong features (the front row of bars) ~30% of features are not represented by any of the novelty neurons, which is close to the prediction of Appendix B.1. The larger the *strength* of the strong features, the larger is the proportion of missed features. For strength = 1.6 (the back row of bars in Fig. 4c,d), almost 80% of the features are not represented, and all the novelty neurons represent almost exclusively the strong features.

Thus, the simulations shown in Figure 4c,d highlight a serious weakness of the combined models: if a group of features is stronger than the rest (e.g., by 60% in Fig. 4c,d), almost all the novelty neurons represent the strong features, and nearly all the other features are discarded. Discarding very weak features may sometimes be profitable to emphasise the most important aspects of stimuli (O'Reilly and Munakata, 2000), but Figure 4c,d shows that the combined models discard not only very weak features, but almost all but the strongest. The combined models discard all features that are weaker than 50% of the strongest features—but such features may contribute very significantly to input patterns and may carry behaviourally relevant information.

The fact that many neurons represent the same strongest features (e.g., in the back row of Figure 4c,d, each feature is represented by five neurons) introduces further significant correlation between the responses of the novelty neurons.

The original model of Sohal and Hasselmo (2000) has a feature extraction mechanism similar to that of the double threshold model, hence it is similarly likely to miss weaker features, as demonstrated in Figure 4d. The mechanisms of feature extraction in the neocortical part of the Norman and O'Reilly (2001) model are more complex than in the combined competitive model analysed here (which is a simplified version of the neocortical part of the original Norman and O'Reilly, 2001 model), but it is likely that the original model will also miss many features. Furthermore, the original Norman and O'Reilly (2001) model includes the mechanism of weight contrast enhancement, i.e., increasing the largest weights and suppressing the smallest, in addition to weight modifications due to Hebbian learning. Norman and O'Reilly (2001)

state that the weight contrast enhancement further increases the network's tendency to focus on the strongest features and to ignore the weaker ones, hence this mechanism will further increase the tendency of the original model to miss weaker features.

To summarise, in the published combined models the neurons "choose" the feature to represent independently from one another, hence there is a probability that many features will be missed. If all the features are equally strongly represented, redundancy may be used to avoid missing features; however, if some features are stronger than others, the neurons have a tendency to represent stronger features and to ignore weaker ones. Hence such networks tend to represent only the strongest features, ignoring (even slightly) weaker features. The fact that many novelty neurons represent the same strongest features in such a case introduces large correlations, hence a very significant reduction in the capacity for familiarity discrimination (see below).

Correlation Between Responses of Real Perirhinal Neurons

The capacities of the models have been analysed assuming that responses of both novelty neurons as well as inputs are uncorrelated (see above, Capacity for Uncorrelated Input Patterns). However, this is likely to be an oversimplification of the real situation in the perirhinal cortex. It will be shown that the existence of correlation between responses can have a major deleterious effect on capacity (see below, Capacity for Correlated Input Patterns). In the present discussion, evidence that perirhinal neuronal responses are correlated is presented.

Erickson et al. (2000) recorded the responses of 169 pairs of distant visually responsive perirhinal neurons (recorded from two electrodes separated by 0.5–8 mm) of behaving monkeys. For each pair of neurons they recorded responses during presentation of 16 or 24 different stimuli, and each stimulus was presented an average of 71 times. For each pair of neurons, they estimated the correlation between the mean responses of the neurons to these stimuli. They obtained a distribution of these estimated correlations with mean $\hat{r}\mu \approx 0.05$ and standard deviation $\hat{r}\sigma \approx 0.313$.

It is possible to show that it is very unlikely that such a distribution of estimated correlations was obtained by chance, i.e., it is very unlikely that the mean responses of the recorded neurons were in reality uncorrelated (note that the correlation estimated from a relatively few pairs of random numbers drawn from uncorrelated distributions is usually different from 0). Thus, the estimated correlation between 16 pairs of uncorrelated numbers is expected to have mean $r_0\mu = 0$ and standard deviation $r_0\sigma \approx 0.258$ (as found numerically by estimating correlations between 16 random numbers 10^7 times). Both mean $\hat{r}\mu$ and standard deviation $\hat{r}\sigma$ observed by Erickson et al. (2000) are significantly larger than the values expected by chance ($\hat{r}\mu \neq r_0\mu$ with $P < 0.05$; $\hat{r}\sigma \neq r_0\sigma$ with $P < 0.0001$). Appendix C.2 uses Monte Carlo methods to estimate the most likely real underlying distribution of the correlations. Hence the observations of Erickson et al. (2000) strongly suggest that it is very unlikely that mean responses of all distant perirhinal neurons are uncorrelated. They rather suggest that some pairs of neurons have positive correlations between their mean responses for different stimuli and some have negative.

Furthermore, Gawne and Richmond (1993) and Erickson et al. (2000) recorded also from pairs of neighbouring visually responsive neurons in inferior temporal cortex (i.e., recorded from a single electrode). These correlations between mean responses of adjacent perirhinal neurons were even greater (i.e., had larger mean and variance) than for the nonadjacent perirhinal neuron pairs. In estimations of the capacity of the familiarity discrimination network in the human perirhinal cortex, the smaller estimate of the correlations between responses of perirhinal neurons is used (ignoring that correlations between adjacent pairs may be even larger) to obtain upper limits of the capacities that can be achieved by various models (see later, Estimation of Capacity of Large Networks). However, simulations will be performed for values of mean correlation covering a range of 0–1, thereby including not only such larger values of mean correlation, but all values likely to exist in the perirhinal cortex (see below, Capacity for Correlated Input Patterns).

Capacity for Correlated Input Patterns

In the followng discussion, we investigate by how much correlation between the responses of different neurons decreases the capacity of the different familiarity discrimination networks. We describe the results of calculations and simulations, first for the Hebbian model, then for the combined models and finally for the anti-Hebbian model. At the end we give an intuitive explanation of the differences in capacities achieved by the various models.

For simplicity of explanation, the capacity was tested using very simple binary patterns. Furthermore, sparse coding was not assumed, i.e., the probability of each input neuron being active was 50%. This simplification may be made as Bogacz and Brown (2002) have shown that the sparseness of coding does not have a great influence on the capacity of familiarity discrimination networks.

These simple binary patterns were generated such that (the modulus of) the correlation between each pair of input neurons was constant. The patterns were generated in the following way. At the beginning of a simulation a binary template pattern x^{temp} was generated randomly. All the patterns x were biased towards x^{temp}, such that the probability of $x = x^{temp}$ equalled $\frac{1}{2} + \frac{1}{2}b$, where b is the parameter that controls bias. For example, for 10 inputs, the template pattern may look like: $x^{temp} = +--+-++--+$, where + denotes that the corresponding input is active, and − that is inactive. If the bias is, for example, equal to 0.6, then on average two inputs in the patterns are different from the template (because the probability of $x = x^{temp}$ equals $\frac{1}{2} + \frac{1}{2}$ of $0.6 = 0.8$), so a sample pattern used in the simulation may look like: $x = ---+-++++-+$. In addition, to keep the level of activity constant across the neurons, the template was inverted, i.e., each bit in the template was switched ($x^{temp} \leftarrow -x^{temp}$) at random moments in time. For patterns generated in this way, the correlation r_{ij} between a pair of inputs was equal to b^2 or $-b^2$.

Analysing capacity using more realistic input patterns (derived in the section, Feature Extraction by Combined Models) is com-

plex and also produces results that are qualitatively the same as those about to be reported (for a detailed discussion, the reader is referred to Bogacz (2001).

In the simulations described in this section, we use the patterns with different levels of correlations between input neurons. The correlations between the responses of the novelty neurons depend on the correlations between the inputs and the learning rule of a particular model. Hence the level of correlations between the responses of novelty neurons is not a free parameter that we modify ourselves, but rather in each of the models it is determined by the bias parameter b of the input patterns. In particular, the Hebbian model assumes for simplicity of analysis that the number of active novelty neurons in the network is limited by the connections with high synaptic weights, hence the correlations between activities of the novelty neurons are the same as between the inputs. If the combined models were extracting independent features, the responses of the novelty neurons would be uncorrelated. However, in currently published combined models (as shown earlier, in the section, Feature Extraction by Combined Models), groups of neurons often represent the same feature, hence their responses are correlated and the magnitudes of correlations between the novelty neurons in the currently published combined models also depend on b.

Appendix C calculates the capacity for the Hebbian model for the case in which activities of the input neurons are correlated. The analytic expression for capacity is found in Appendix C: it shows that the capacity depends on the average correlations between pairs of inputs. Figure 5a compares the theoretical predictions with the results of simulations and shows that the network capacity decreases very markedly even when the correlation is very small. For example, for a network of 200 neurons, the capacity is less than one fifth of that for uncorrelated patterns when $|r_{ij}| = 0.04$ (a value similar to the mean correlation between distant pairs of neurons found by Erickson et al., 2000; on the x-axis in Fig. 5a, see: $\mathrm{sqrt}(|r_{ij}|) = 0.2$).

Appendix C also shows that the impact of correlation on the Hebbian model's capacity is reduced when the connections between input neurons and novelty neurons are sparse. Thus, when neuronal activities are correlated, the capacity per synapse is greater for sparse connections (note that here we are considering sparse connections, not sparse coding). Figure 5b compares the theoretical predictions with the results of simulations for a network of 100 neurons. The results of the simulations show that a network in which each novelty neuron receives connections from 40% of the network's inputs achieves approximately one-half the capacity of the fully connected network for uncorrelated patterns (i.e., $|r_{ij}| = 0$). But for $|r_{ij}| = 0.04$ (on the x-axis in Fig. 5b, see: $\mathrm{sqrt}(|r_{ij}|) = 0.2$) the capacity of the sparsely connected network is reduced less than that of the fully connected network so that the two capacities now differ by only 15%, and for $|r_{ij}| = 0.09$ (on the x-axis in Fig. 5b see: $\mathrm{sqrt}(|r_{ij}|) = 0.3$) they are equal.

Appendix D gives the upper bound for capacity for correlated input patterns of the class of single layer combined models with Hebbian learning. It shows that correlation in the input patterns decreases the capacity of the combined models according to an equation similar to that for the Hebbian model. However, if a combined network were to complete feature extraction (i.e., with statistically independent activities of the novelty neurons), correlation in the input patterns would reduce the capacity of the combined models less than for the Hebbian model. However, it has been demonstrated that the currently published combined models do not extract statistically independent features (as shown earlier, in the section, Feature Extraction by Combined Models). Therefore, they do not achieve this upper limit. Furthermore, Figure 5c shows that there is a decrease in capacity for both the combined competitive and the double threshold models that is even greater than that for the Hebbian model. For example, for $|r_{ij}| = 0.04$ (on the x-axis in Fig. 5c see: $\mathrm{sqrt}(|r_{ij}|) = 0.2$) and 200 neurons, the capacity is less than one sixth of that for uncorrelated input patterns.

Appendix E calculates the capacity of the anti-Hebbian model for correlated input patterns, and shows that the anti-Hebbian model is very robust to the correlation between the responses of the input neurons. This prediction is consistent with the results of simulations presented in Figure 5d, which shows that the correlation between responses of input neurons reduces the capacity of the anti-Hebbian model much less than other models. For example, for $|r_{ij}| = 0.04$ (on the x-axis in Fig. 5d see: $\mathrm{sqrt}(|r_{ij}|) = 0.2$) and 200 neurons, the capacity is almost 90% of that for uncorrelated input patterns.

Furthermore, in the familiarity discrimination networks based on Hebbian learning, the influence of the correlation between responses of input neurons on capacity increases when the size of the network grows. For example, for the Hebbian model, the capacity for $|r_{ij}| = 0.01$ (on the x-axis in Fig. 5a see: $\mathrm{sqrt}(|r_{ij}|) = 0.1$) is ~90% of that for uncorrelated input patterns for $N = 100$ neurons, ~70% for $N = 200$ neurons, and ~50% for $N = 300$ neurons. By contrast, in the anti-Hebbian model, the influence of the correlation between responses of input neurons on capacity decreases when the size of the network grows. For example, for the anti-Hebbian model, the capacity for $|r_{ij}| = 0.25$ (on the x-axis in Fig. 5d see: $\mathrm{sqrt}(|r_{ij}|) = 0.5$) is ~42% of that for uncorrelated input patterns for $N = 100$ neurons, ~61% for $N = 200$ neurons, and ~66% for $N = 300$ neurons. Hence for large networks, the anti-Hebbian model achieves a capacity much greater than any of the networks based on Hebbian learning when there are even very small correlations between the responses of the input neurons.

This difference in capacities of the combined models and anti-Hebbian model may be explained intuitively by the fact that the combined models extract features, hence they focus on the elements common to all the input patterns (i.e., features), while the anti-Hebbian model ignores the features and focuses on the elements characteristic for individual patterns.

This principle is illustrated in Figure 6, which compares the weight changes of the combined competitive and the anti-Hebbian models during presentations of features repeating in the stimuli. In the scenario of Figure 6, four patterns are presented, which include repeating features: the same feature is present in the first and third patterns and another feature is present in the second and fourth. Let us imagine that Figure 6 shows only a part of the input patterns and a part of the familiarity discrimination network, i.e., each of the input stimuli is encoded by the four inputs shown in Figure 6,

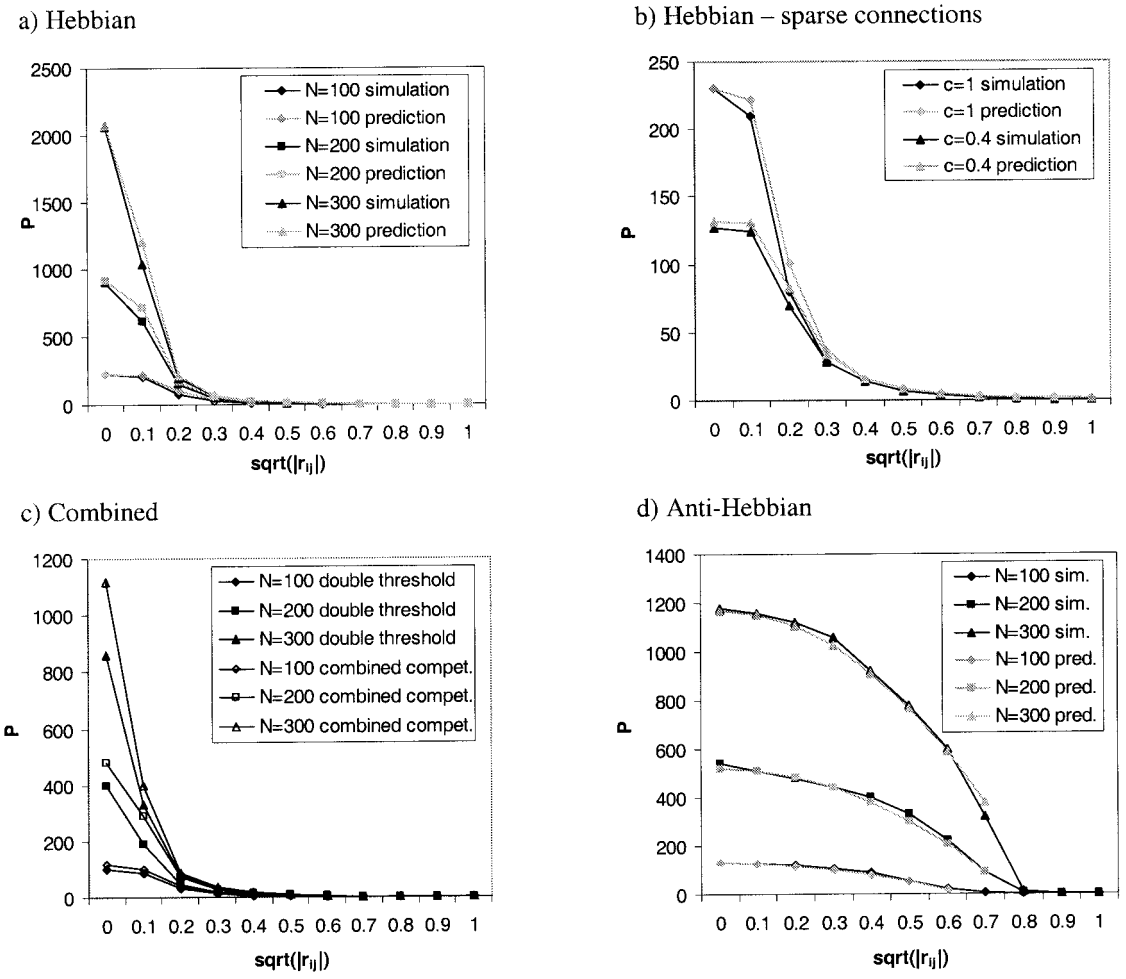

FIGURE 5. Capacities (number of stored patterns, P) of familiarity discrimination networks for correlated (square root of $|r_{ij}|$) patterns. a: Capacity of the fully connected Hebbian model with correlated novelty neurons and correlated inputs. The theoretical predictions (a and b) are slightly higher than the results of simulations due to the approximations made in Appendix C.1 (discarding cases 2 and 3 from Table 1). b: Capacity of the Hebbian model of N = 100 neurons with diluted connectivity for a fully connected network (series c = 1) and a network with 40% of the connections present (series c = 0.4). c: Simulated capacity of combined models. d: Simulated capacity of the anti-Hebbian model. Note that the influence of correlation on the capacity of the combined models (c) is similar to that for Hebbian model (a), and much stronger than on capacity of the anti-Hebbian model (d). Methods of calculating capacity as in Bogacz et al. (1999, 2001a). For each network and for each number of neurons N, the familiarity discrimination error was estimated for different numbers of stored patterns P, and the capacity P_{max} was taken as the maximum number of stored patterns P, for which the error rate was ≤1%. For given N and P, the discrimination error was estimated in the following way. During each test, P patterns were presented to the network, and then accuracy was tested on all the presented patterns (i.e., from the list) and equal number of novel patterns (i.e., patterns not from the list, but also generated in the same way as for stored patterns). These tests were repeated until the network had been tested with 5,000 previously presented patterns and 5,000 random (novel) patterns, e.g., for P = 100, the tests were repeated 50 times. The average accuracy over the tests is taken as a result. To illustrate the precision of the simulation process, for one data point (100 neurons), the capacity was estimated 10 times using the above method. The standard deviation of the estimated capacities was ±5.4 (i.e., about 2.5% of the mean). The capacity was tested on randomly generated patterns, such that the modulus of the correlation between any two inputs was constant: $|r_{ij}|$ = const. (for details of pattern generation, see Capacity for Correlated Input Patterns), and the square root of $|r_{ij}|$ is shown on the x-axes. The square root of $|r_{ij}|$ is equal to the bias towards template b (see Capacity for Correlated Input Patterns) and it is plotted to emphasise the rapid drop of capacity even for very small values of correlation $|r_{ij}|$. The y-axes show capacity P (note that the scales are different).

and also by a number of other inputs that are unique for each of the stimuli but are not shown. The combined competitive model extracted the features, and the weights of each neuron represented one feature. By contrast, the anti-Hebbian model ignored the repeating features, and the repeated presentation of the features did not change the weights in the network. This example illustrates that the combined models learn the features common to all the stimuli, and since the features occur in many stimuli, they are represented in the weights of the novelty neuron with much higher magnitude than the elements characteristic for individual input patterns. By contrast, the anti-Hebbian model ignores the repeating features—they are not represented in the weights with large

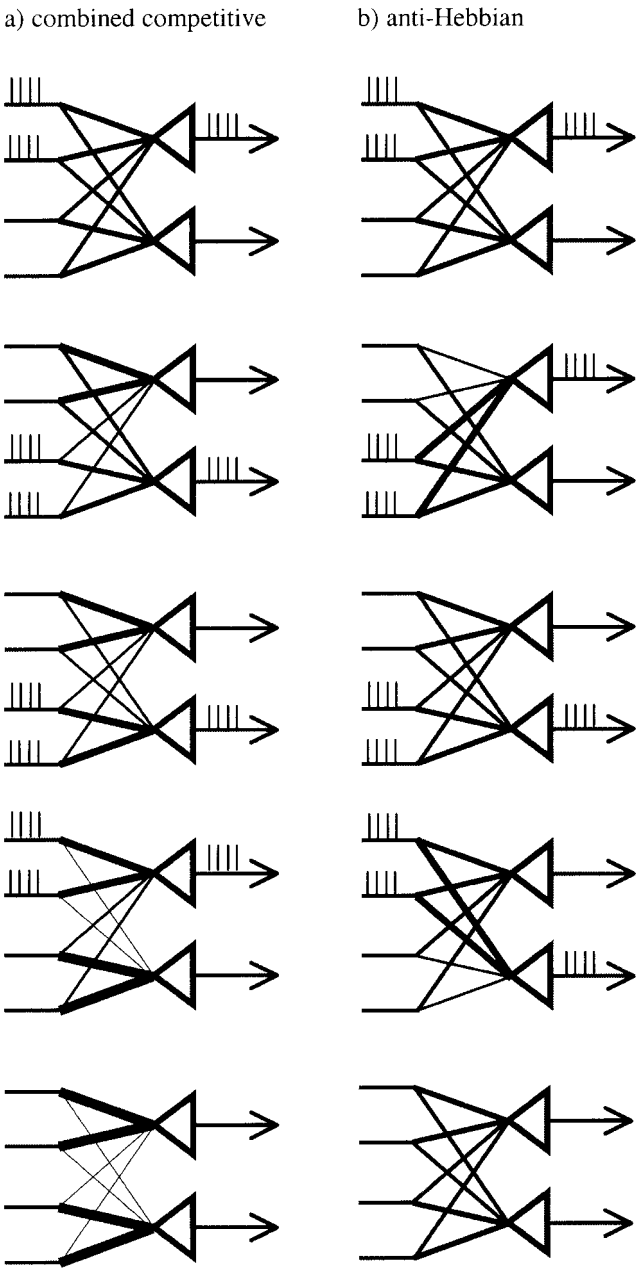

FIGURE 6. Comparison of example weight modifications of (a) combined competitive and (b) anti-Hebbian model. Notation as in Fig. 1. The upper row of panels shows initial values of the weights and is the same for both networks. Four stimuli are presented to each network and the following four rows of panels show the weights resulting from the weight modification according to the stimulus in the panel above. The four patterns include two repeating features. The first feature is represented by the activity of the two upper neurons, and the second feature by the activity of the two lower neurons. Note that after stimulus presentations the combined competitive model learned the features (i.e., the weights of each neuron represent one of the features), while the anti-Hebbian model ignored the features (i.e., the weights after stimulus presentations are exactly the same as at the beginning).

magnitude—and hence the weights can represent elements characteristic for individual patterns. Therefore, whenever there exists even a small correlation between the activities of the input neurons (i.e., input patterns share some features), the anti-Hebbian network can judge the familiarity of many more individual stimuli than the combined models. Assuming that feature extraction networks have been used to achieve perceptual identification of visual stimuli on the way to and within perirhinal cortex, it is perhaps not surprising that a network based on a different learning rule is then more efficient in performing a different type of task (familiarity discrimination).

To summarise, this section shows that any correlation between responses of input neurons reduces the capacity of the Hebbian model very substantially. The theoretical upper limit of capacity of the combined models is slightly above the capacity of the Hebbian model, but since the currently published combined models do not extract statistically independent features, simulations indicate that they achieve capacities even lower than the Hebbian model when the activities of inputs are correlated. By contrast, the anti-Hebbian model achieves a much higher capacity than the models based on Hebbian learning when the activities of inputs are correlated. This comes from the fact that the combined models learn features common to all the stimuli, while the anti-Hebbian model learns elements characteristic for individual input patterns, and hence can recognise many more individual stimuli than the combined networks.

Estimation of Capacity of Large Networks

Figure 7 shows the predictions of the capacity of putative networks of novelty neurons in the human perirhinal cortex, for a network of 4×10^6 novelty neurons, each receiving either 10,000 (Fig. 7a) or 1,000 (Fig. 7b) input connections. The curves with filled diamonds show the capacity of the Hebbian model. The curves are based on the predictions of capacity calculated in Appendix C and verified in the simulations of Figure 5a,b. The curves with open squares show the capacity of the anti-Hebbian model. They are based on the predictions of capacity calculated in Appendix E and verified in the simulations of Figure 5d. These curves are constant, as the correlations considered in Figure 7 are too small to have any effect on the capacity of so large network (such correlations have already little and diminishing effect on the capacity of the relatively small networks considered in Fig. 5d). The expected capacities of the combined competitive and double threshold models are not illustrated, but it could be extrapolated from simulations that they would lie below the capacity of the Hebbian model (i.e., below the lines with filled diamonds) (see above, Capacity for Correlated Input Patterns).

Figure 7 shows that the average correlation (defined as in the legend of Fig. 7) must be extremely small (i.e., <0.00001), for the capacity of the Hebbian model to be the same as for the uncorrelated patterns. When the correlation is larger (i.e., >0.0001), then the capacity of the Hebbian model decreases by 1.5 orders of magnitude ($\times\sim 30$) when the correlation increases by one order of magnitude ($\times 10$). This power law is also evident from the equations for capacity in Appendices C and D.

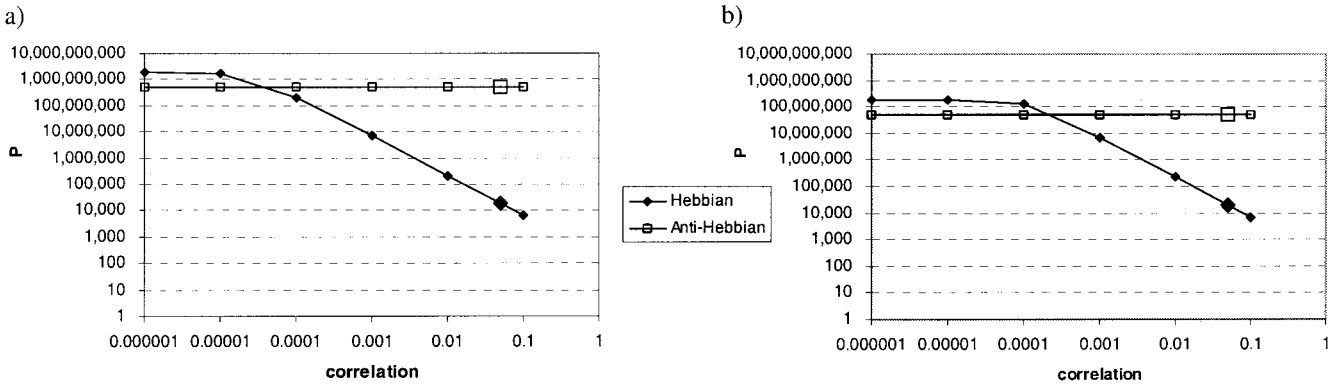

FIGURE 7. Predictions of capacity (P) for a network of novelty neurons in the human perirhinal cortex. Predictions are made for a network of 4×10^6 novelty neurons, each receiving 10,000 connections from inputs (a) or 1,000 connections (b); *y*-axis shows the capacity; x-axis shows the correlation in the input patterns defined as a cube root of parameter r^3 defined in Appendix C (note that x-axis shows correlation rather than square root of correlation as used in Fig. 5). This measure of correlation is closely related to the average absolute correlation between all pairs of inputs. The two measures are exactly equal (i.e., correlation = $|r_{ij}|$) when the values of correlation between pairs of inputs are the same for all pairs of inputs (as they are in the simulations of Fig. 5). Both axes have a logarithmic scale. Larger symbols show capacity for correlation equal to 0.05. The value of 0.05 is the average correlation between responses of distant perirhinal neurons measured by Erickson et al. (2000), and it is also very close to the cube root of parameter r^3 calculated for the values of correlations between responses of distant perirhinal neurons (see Appendix C.2).

Comparing Figure 7a and 7b shows that for larger correlations (>0.001) the capacity of the Hebbian model when each novelty neuron receives 10,000 inputs is the same as that when each novelty neuron receives 1,000 inputs. Hence, above a certain average value of the correlation, there is no advantage in increasing the number of synapses of each neuron above a particular number. Thus, the magnitude of the correlation in a network such as that of the Hebbian model may determine the optimal number of synapses per neuron.

Figure 7 shows that for larger correlations between responses of input neurons the anti-Hebbian model achieves a capacity much larger than the networks based on Hebbian learning. In particular, the larger squares and diamonds show the capacity for a correlation of 0.05, close to the value measured by Erickson et al. (2000). For this value of correlation, the anti-Hebbian model achieves capacity 1,000–10,000 times larger than the capacity of the networks based on Hebbian learning.

In addition, the capacity shown in Figure 7 is an overestimation of the capacity of the human perirhinal network that may be achieved by the various models, because the capacity of familiarity discrimination networks is decreased by noise (Bogacz, 2001). Moreover, the correlation of 0.05 is an underestimation of the mean correlation between responses of perirhinal neurons, because it does not take into consideration the larger correlations found between responses of adjacent perirhinal neurons (see above, Correlation Between Responses of Real Perirhinal Neurons). It follows that if we assume that the correlation between responses of input neurons is similar to that observed by Erickson et al. (2000), the currently published models based on Hebbian learning cannot achieve a capacity sufficient to explain human recognition capabilities. However, this conclusion does not apply to the anti-Hebbian model; its capacity remains sufficiently high in the presence of correlated activity and noise.

However, it should be noted that there is no evidence that the correlations observed by Erickson et al. (2000) were measured between pairs of neurons such that both were providing input to the familiarity discrimination network. There are many different functional types of neurons in the perirhinal cortex (Fahy et al., 1993; Xiang and Brown, 1998; Brown and Xiang, 1998; Murray and Bussey, 1999). It will be very important to establish which of them are correlated with which other types, as well as the mean correlations between the various groups of neurons (see below, Suggested Experiments), because it will establish whether the models based on Hebbian learning could explain human recognition abilities.

Detecting Unusual Stimuli

If we assume that correlation in the responses of input neurons arises from regularities in experienced stimuli, unusual stimuli should result in a pattern of activity of input neurons that differs from that usually observed. The problem of finding unusual patterns that belong to a different distribution from that of typical patterns is often called novelty detection, and many neural network algorithms for it have been proposed (e.g., Bishop, 1994; Roberts and Tarassenko, 1995; Parra et al., 1996). All the models reviewed in this study can also detect unusual stimuli. The probability of correct detection of an unusual stimulus by the Hebbian model is calculated in Appendix F and presented in Figure 8a, which shows that the larger the correlation between inputs for typical stimuli and the bigger the size of the network, the smaller is the probability of error. The currently published combined models have a similar ability to detect unusual stimuli (Fig. 8b). However, the anti-Hebbian model detects unusual stimuli with much less accuracy (Fig. 8b); this could be expected as the anti-Hebbian model ignores the features common to usual stimuli (see above, Capacity for Correlated Input Patterns).

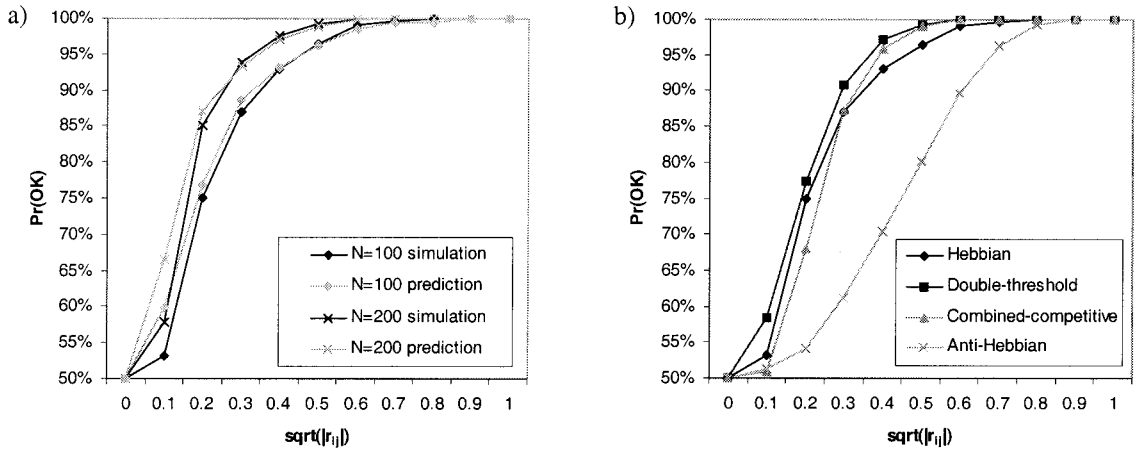

FIGURE 8. Error of discrimination between normal and unusual stimuli. For each network, 1,000 typical correlated patterns were presented and the probability of error was approximated by taking the error of discrimination for 5,000 typical patterns (i.e., patterns coming from the same distribution as the previous 1,000), and 5,000 unusual patterns (i.e., random). Patterns as in Fig. 5; x-axis shows the square root of correlations between inputs, as in Fig. 5, y-axis shows the proportion of correct discriminations. a) Comparison of error of the Hebbian model obtained in simulations with predictions of Appendix F. b) Comparison of errors of different networks obtained in simulations for N = 100 neurons. Note that networks based on Hebbian learning detect unusual stimuli with greater accuracy than the anti-Hebbian model.

The ability to detect unusual stimuli or features is likely to be very useful during feature extraction as novel unusual stimuli may, for example, require greater amendment to their representation. Hence, it seems quite possible that an ability to detect such novelty may be a general characteristic of networks extracting features in the earlier stages of visual processing and during the early stages of an individual's development.

DISCUSSION

It has already been established that when the responses of neurons providing input to a familiarity discrimination network are correlated (as is indicated by experimental data), specialised networks achieve higher capacity than combined networks, because the currently published combined models do not extract statistically independent features (see the section, Comparison of the Models' Efficiencies). Furthermore, the specialised anti-Hebbian model may achieve capacity a few orders of magnitude higher than the specialised Hebbian and combined models, because the latter models learn features common to multiple stimuli, while the anti-Hebbian model learns elements characteristic for individual stimuli. Hence, the combined models as so far published are critically less efficient than the specialised model based on anti-Hebbian learning.

Attempts to combine feature extraction with familiarity discrimination encounter a number of problems that are avoided when the processes are separated. Thus, it has been shown that the currently published combined models have a tendency to ignore (i.e., not extract) many of the features present in the input patterns (see above, Feature Extraction by Combined Models). If all the features are equally strong, there need to be 4.6 more novelty neurons than features to miss only 1% of features. However, if some features are stronger than others, the great majority of the novelty neurons represent the stronger features, while (even slightly) weaker features are ignored. It is likely that features present in natural visual stimuli are not equally strong. This critically limits the ability of the currently published combined models efficiently to extract features. Furthermore, the fact that many novelty neurons represent the same strongest features introduces correlations that will very significantly reduce capacity. It remains possible that a combined familiarity discrimination network with an architecture designed to ensure extraction of independent features may achieve the high theoretical limit of capacity. However, as yet it has not been possible either to design or to prove that such networks exist.

There is evidence that representation of visual stimuli in the inferior temporal cortex is modified by experience in adult monkeys. Kobatake et al. (1998) trained monkeys to distinguish between the elements of a set of similar geometrical figures. These investigators found that the responses of inferior temporal neurons before training were very similar for the different figures. After training the responses differed more greatly between the different figures, which presumably allowed monkeys to discriminate more accurately between them (Kobatake et al., 1998). However, one may argue that learning representations is a long-lasting process requiring gradual small synaptic changes after repeated presentations of stimuli. This process may be expected to be most intensive in development and childhood. By contrast, recognition memory is a process requiring rapid changes in synapses even after a single stimulus presentation (to allow recognition of its prior occurrence during a subsequent presentation), which occurs during both childhood and adulthood. In other words, a network extracting independent features may require a different (i.e., smaller) magnitude of synaptic weight modifications after a single stimulus pre-

sentation than a network discriminating familiarity. It remains to be established whether the same magnitude of weight modification after a single stimulus presentation may allow a combined network to learn a sufficiently stable representation and at the same time recognise a stimulus as familiar on its second presentation.

In particular, most of the novel stimuli we are normally exposed to are merely new combinations of known features, e.g., the face of a person we meet for the first time is a combination of previously known feature types (e.g., eyes, lips, nose) similar to those seen in other people. Hence, such novel stimuli do not require large alterations of representation (i.e., feature extraction), but they do require a large modification of the weights of the novelty neurons in order to recognise the stimuli during their future occurrences. In contrast, a large alteration in representation is required when we are exposed to unusual stimuli with features we have never seen before, e.g., if surrounded by extraterrestrial aliens or, as in the Kobatake et al.'s (1998) experiment, where discrimination among a set of similar stimuli is behaviourally important. Hence, feature extraction networks should significantly alter representations only when they detect unusual stimuli. Feature extraction networks have the ability to detect unusual stimuli independent of their capacity for familiarity discrimination (see above, Detecting Unusual Stimuli). So the fact that most novel stimuli we perceive seem not to require large alterations of the weights of the feature extraction network (as they are combinations of known features), but do require significant alterations of the weights of the familiarity discrimination network, also suggests that feature extraction and familiarity discrimination should be performed by separate networks. An interesting possibility is that if the principles of the combined models are used in more posterior stimulus categorisation networks, the reduction in the number of neurons that respond to repeated stimuli in such models might provide a neuronal basis for the facilitation of performance (increased perceptual fluency) seen in repetition priming.

We have not presented an analysis of the capacity of the Hebbian inhibitory model (Fig. 1c). However, because the operation of the Hebbian inhibitory model is equivalent to the operation of the anti-Hebbian model (see above, Independent Responses of Novelty Neurons), the Hebbian-inhibitory model may potentially also achieve a similar very high efficiency with respect to the number of synapses used for recognition memory, similar to that of the anti-Hebbian model (i.e., a capacity proportional to the number of modifiable synapses). In a comparison of the consistency of this model with the results of experimental observations (see below, Synaptic Plasticity), it is pointed out that its capacity is necessarily lower than that of the anti-Hebbian model as the potential number of modifiable synapses (being inhibitory rather than excitatory) is lower (Thompson et al., 2001).

To summarise, it is established that when the responses of neurons providing input to a familiarity discrimination network are correlated, the Hebbian model has a much lower capacity than the anti-Hebbian model (see the section, Comparison of the Models' Efficiencies). For the combined models, it is necessary to assume that activities of the inputs may be correlated, as by hypothesis these networks have not yet completed feature extraction. The theoretical upper limit of capacity of the combined models is slightly above the capacity of the Hebbian model, but the currently published combined models do not extract statistically independent features. This produces a marked reduction in their capacity due to three effects: (1) correlations between the activities of novelty neurons; (2) the redundancy required in the number of novelty neurons in order to extract most of the equally strong features (which further reduces capacity at least 4.6 times); and (3) the fact that the novelty neurons represent only the strongest features while ignoring the weaker ones may reduce the capacity many times depending on the disproportions in strength of the features. Furthermore, if a combined network is to be considered as a viable solution, it needs to be demonstrated that the same magnitude of weight modification after a single stimulus presentation will allow the combined network to learn a sufficiently stable representation at the same time as achieving single exposure learning. Even without these considerations, if the correlations between the input neurons have values similar to those measured between the responses of distant perirhinal neurons by Erickson et al. (2000), then familiarity discrimination networks of the size of that in the human perirhinal cortex working according to the models based on Hebbian learning (i.e., Hebbian and combined models) would achieve a capacity a few orders of magnitude lower than a network working according to the anti-Hebbian model. Accordingly, the arguments concerning efficiency favour there being a separate, specialised network for familiarity discrimination in perirhinal cortex.

CONSISTENCY OF THE MODELS WITH EXPERIMENTAL OBSERVATIONS

This section compares the consistency of the models with what is known of possible types of synaptic plasticity and of the responses of perirhinal neurons. In addition, psychological features and other experimental results that have a bearing on whether or not there exists a network in the perirhinal cortex specialised only for familiarity discrimination are described, and further experiments are suggested.

Psychological Features

Norman and O'Reilly (2001) compared how networks performing familiarity discrimination and recall behave in recognition memory tasks. Their full model is much more complex than the simple combined competitive model analysed in this study and includes more sophisticated models of the neurons, their dynamics, and synaptic plasticity. Hence their model explains in an elegant way various results of psychological experiments, for example: the shape of ROC curves, the effect of list strength, and the effects of lesions to the medial temporal lobe (Norman and O'Reilly, 2001).

Although Norman and O'Reilly (2001) proposed that the network in the inferior temporal cortex combines feature extraction and familiarity discrimination, they performed all their simulations for a very small number of stimuli (e.g., 20). Thus, rather than establishing that their model could extract features, these

investigators focused on the consistency of the model's behaviour in recognition memory tests with the results of psychological experiments. Norman and O'Reilly (2001) used patterns with correlated responses for the inputs in their simulations, as this correlation may be expected often to exist in patterns representing stimuli used in psychological experiments because these stimuli often share many features (e.g., all of them are words). The Norman and O'Reilly (2001) model could also work as a specialised network with an efficiency similar to that of the Hebbian model (the efficiency of the combined competitive model for uncorrelated input patterns has been analysed by Bogacz and Brown, 2002). Hence the differences between the properties of the familiarity discrimination and recollection networks shown by Norman and O'Reilly (2001) are valid no matter whether it is assumed that there exists a specialised or a combined network in the perirhinal cortex. Thus, Norman and O'Reilly's (2001) analysis is focused on establishing which additional mechanisms are necessary to observe not only efficiency but also the shortcomings and mistakes made by human subjects in recognition memory tasks (K. Norman, personal communication). By contrast, here we analyse which basic network mechanisms allow the capacity to be large enough to store all of the patterns that we encounter (and can discriminate between) over a lifetime.

Bogacz et al. (2001a) demonstrated the Hebbian model shows a false memory effect, but it is plausible that the behaviour of all the models would be consistent with this observation, because of the similarity in their computational bases.

Synaptic Plasticity

The primary synaptic mechanism in the anti-Hebbian model is homosynaptic LTD (or equivalent mechanism). The model also assumes an increase in the strength of synaptic connections from inactive inputs (Fig. 1a). Homosynaptic LTP and LTD have been demonstrated in the perirhinal cortex (Bilkey, 1996; Ziakopoulos et al., 1999; Cho et al., 2000). Furthermore, homosynaptic LTD is easier to produce in perirhinal slices if the postsynaptic neuron is depolarised than if it is hyperpolarised (Cho et al., 2000). This pattern is consistent with the anti-Hebbian model because the weights of active novelty neurons should be modified, but the weights of inactive should not. However, these brain slice experiments indicate that homosynaptic LTD is produced in perirhinal slice by low-frequency stimulation (1 Hz), while high-frequency stimulation produces LTP (Cho et al., 2000; Ziakopoulos et al., 1999). This appears to be inconsistent with the anti-Hebbian model which assumes that the weights from more active inputs should be decreased while the weights from less active increased. Because the relationship between slice experiments and plasticity in the functioning brain is still unclear, the role of perirhinal LTD in familiarity discrimination needs to be clarified. In particular, it is important to establish the type of synaptic change produced in perirhinal neurons studied in vitro when stimulation patterns mimic those indicated by perirhinal neuronal responses in vivo.

Experimentally, LTP in the perirhinal cortex is induced by activating both presynaptic and postsynaptic neurons, while in this model the compensatory increase in synaptic strength of synapses not undergoing LTD should be produced by heterosynaptic LTP at inactive synapses. Such an increase has not been observed at synapses of inputs from one side of the perirhinal cortex (e.g., the temporal side), after producing LTD at synapses of inputs from the other side (e.g., the entorhinal) (K. Cho, personal communication), and has not been reported for any other brain region. However, to our knowledge, this kind of synaptic weight increase consequent upon LTD within the same pathway (e.g., the same side) has not yet been sought in the perirhinal cortex. Nevertheless, such a mechanism of heterosynaptic LTP may seem somewhat implausible. However, the necessary effect could be achieved simply by increasing the strength of the other synapses of a given neuron, after homosynaptic LTD at some of its synapses, in such a way as to maintain the neuron's excitability, thereby removing the requirement for heterosynaptic LTP.

Recently, the homeostatic mechanisms that act to maintain average neuronal activity and thus promote network stability have been reported in cultures and slices of cortical neurons (for review see Turrigiano and Nelson, 2000). The homeostatic mechanisms include the following:

1. *Scaling synapses*: Turrigiano et al. (1998) reported that a decrease in activity of neurons in a culture caused an increase in postsynaptic responsiveness to glutamate agonists
2. *Regulation of neuronal excitability*: Desai et al. (1999) reported that a decrease in activity of neurons in a culture caused a decrease of the threshold for spike generation
3. *Regulation of synapse number*: it has been observed that synaptic density in slices and cultures varies inversely with neuronal activity (Kirov et al., 1999; Kirov and Harris, 1999; Segal, 2001).

Such homeostatic mechanisms could also be employed to maintain network excitability should the principles of the anti-Hebbian model be implemented in the real perirhinal cortex.

The models based on Hebbian learning, i.e., the Hebbian model and all the combined models, rely upon homosynaptic LTP, homosynaptic LTD, and heterosynaptic LTD. The occurrence of heterosynaptic LTD has not been described in the perirhinal cortex, though it has been found elsewhere (Ito, 1989).

The Hebbian inhibitory model assumes the existence of plastic synapses between inhibitory and excitatory neurons in the perirhinal cortex. Such plasticity is known to exist in the cerebellum (Mitoma and Konishi, 1999), but there are no published reports of such plasticity in the perirhinal cortex. Furthermore, in the perirhinal cortex, the number of excitatory synapses is much larger than the number of inhibitory synapses (Thompson et al., 2001). Accordingly, the capacity of the perirhinal cortex for familiarity discrimination would be greatly reduced should it be mediated by the plasticity of only a small fraction of its synapses.

To summarise, all the proposed models assume the existence of one or more synaptic mechanisms they have not yet been identified in the perirhinal cortex. But the most plausible synaptic plasticity is assumed by the models based on Hebbian learning and the least plausible by the Hebbian inhibitory model. The anti-Hebbian model relies on a mechanism of compensatory excitability changes that is not yet proven, although not implausible.

Responses of Perirhinal Neurons

The double threshold model predicts that the neurons which initially are the most active for a novel stimulus should increase their activity even more for subsequent presentations of the stimulus, while neurons initially less active should decrease their activity for subsequent presentations of the stimulus.

Li et al. (1993) sought this effect in the responses of 72 visually responsive anterior inferior temporal, including perirhinal, neurons during repeated presentations of initially novel stimuli. For 24 of the 72 neurons, there was a significant change in response with repetition: in all cases this change was a decrease. Similarly, Xiang and Brown (1998) found only 8 ($<1\%$) of 1122 visually responsive neurons had responses that increased (and none of them strongly) upon stimulus repetition, whereas responses decreased for 446 (40%) of the sample. Furthermore, Li et al. (1993) observed that for a given neuron, the stimuli that elicited a strong initial response (during their first presentations) showed a greater response decrement over their subsequent presentations than did stimuli that were initially less effective (i.e., which elicited a weaker response during their first presentations).

Thus, the experimental observations in perirhinal cortex are opposite to the predictions of the double threshold model. However, an increase in the response of a neuron for a certain stimulus as it becomes familiar has been observed earlier in the visual processing stream, in cortex near the superior temporal sulcus (Rolls et al., 1989). Nevertheless, this area is not involved in long-term familiarity discrimination, because presentation of a novel stimulus influences the neuronal response to a subsequent presentation of this stimulus only when the presentations are separated by less than 7 trials, i.e., $\sim$1 minute (Rolls et al., 1989).

Miller and Desimone (1994) observed that in a delayed matching to sample task when monkeys had been trained that response to a repeated sample stimulus was rewarded while other repeated stimuli were not, some inferior temporal neurons had larger responses to the second presentation of the target stimulus. However, such increases have only been described when an animal has been so trained and have not been established for more general conditions when more than one stimulus must be held in mind at a time. Indeed, even those neurons for which such increases were found tended to show reduced responses to the repeated presentations of incidental stimuli. Hence such response increases must be part of a different mechanism from that for general familiarity discrimination (Brown and Xiang, 1998).

The perirhinal observations reported by Li et al. (1993) are consistent with the anti-Hebbian and the Hebbian inhibitory models. The observations can be also explained within the Hebbian and the combined competitive models if it is assumed that the magnitude of weight modification is higher for stimuli that are classified as novel with higher confidence (as manifested by the higher population activity of the novelty neurons). Bogacz et al. (2000) have shown that this mechanism is easy to implement within a biologically plausible familiarity discrimination network, and it further increases the capacity.

In both the Hebbian and the combined competitive models, the response of novelty neurons is lower for familiar stimuli due to inhibition, but according to these models the response of novelty neurons should be higher for familiar stimuli in the brief initial interval (before suppression by inhibition). However, simulations (Bogacz et al., 2001a; 2001) show this interval may be very brief (e.g., 10 ms) and, due to temporal jitter, the increase in firing rate for familiar stimuli is not readily visible in peristimulus time histograms of either simulated or real neuronal responses.

The Hebbian and the combined competitive models assume that inhibitory neurons should have higher responses for familiar than for novel stimuli. Neurons with such responses have not been found in monkey perirhinal cortex, although large numbers of neurons have been analysed (Fahy et al., 1993; Xiang and Brown, 1998; Li et al., 1993; Sobotka and Ringo, 1994). However, there is in fact no evidence that inhibitory neurons have ever been recorded. Furthermore, neither model requires there to be large numbers of such inhibitory cells.

Note that the Hebbian inhibitory model (in contrast to the Hebbian and the combined competitive models) does not predict increased activity of inhibitory neurons for familiar stimuli—their activity remains the same, but each neuron receives more inhibition after presentation of familiar stimuli due to the synaptic weight modifications. The Hebbian inhibitory model also predicts that at least some inhibitory perirhinal neurons should be stimulus selective. Although a large proportion of perirhinal neurons are stimulus selective (Li et al., 1993, Miller et al., 1993; Xiang and Brown, 1998), it is unknown whether inhibitory neurons are among them.

To summarise, the anti-Hebbian model is fully consistent with the observation of the responses of perirhinal neurons, the predictions of the double threshold model contradicts the observation that initially most active neurons decrease their responses most (Li et al., 1993), while all the other models require verification of their predictions concerning the responses of inhibitory neurons.

Combined or Specialised Network?

The idea of a network combining familiarity discrimination with feature extraction has the advantage of elegance. One could speculate that such networks are present in the perirhinal cortex and in the earlier sensory areas, and the perirhinal cortex discriminates the familiarity of whole stimuli just because this is the first area where the entire stimulus (as opposed to its individual features) is represented (Buckley and Gaffan, 1998; Murray and Bussey, 1999). If this were so, it is plausible that an increase in firing for novel stimuli is not observed in earlier sensory areas because individual features (e.g., lines, colours) are already highly familiar, as they have previously been repeatedly experienced as part of very many different stimuli. Nevertheless, certain findings relating to the perirhinal cortex favour the possible existence of a specialised familiarity discrimination network.

First, the perirhinal cortex differs anatomically from earlier sensory regions, suggesting its possible functional specialisation. True perirhinal cortex, Brodmann's area 35, does not have a layer IV (Suzuki and Amaral, 1993; Burwell, 2000). In addition, the perirhinal cortex does not have a columnar organisation, in contradistinction to neocortical areas (Suzuki and Amaral, 1993). Fur-

ther, the inputs to the perirhinal cortex from other areas do not have a clustered topographic organisation (Suzuki and Amaral, 1993). Moreover, the adjacent area 36 has extensive intrinsic connections, such that any location is connected to the entire subregion (Burwell, 2000). In these respects, the connectivity of the perirhinal cortex is distributed rather than clustered, and so has similarities to that of the hippocampus.

Second, experimental observations suggest that the neurons involved in familiarity discrimination in the perirhinal cortex create a separate network. Xiang and Brown (1999) recorded responses of perirhinal neurons while monkeys were performing a serial recognition task, and then recorded the responses of the same neurons while monkeys were performing a conditional visual discrimination task. They showed that perirhinal neurons that responded differentially in the serial recognition task (i.e., during familiarity discrimination) constituted a population separate from the perirhinal neurons that respond differentially in the conditional visual discrimination task.

Third, Baxter and Murray (2001) compared the performance of four monkeys in object identification and recognition memory tasks both before and after a neurotoxic lesion to rhinal cortex (perirhinal and entorhinal). These investigators observed that the effects of the lesion on discrimination learning were not reliably correlated with the deficits in recognition memory (Baxter and Murray, 2001). Hence, the magnitude of the discrimination learning impairment did not predict the magnitude of the recognition memory impairment. The absence of an association between object discrimination impairment and object recognition impairment suggest that these memory functions are mediated by separate mechanisms within the rhinal cortex (Baxter and Murray, 2001).

Fourth, an interesting indication of the existence of a separate familiarity discrimination network is the specialisation of neurons involved in familiarity discrimination. Xiang and Brown (1998) recorded the responses of perirhinal neurons after presentation of four types of stimuli:

Novel-first (N1): stimuli seen by the monkey for the first time

Novel-second (N2): unfamiliar stimuli presented for the second time during the recording session

Familiar first (F1): stimuli familiar (well known) to the monkey presented for the first time during the day of recording

Familiar second (F2): familiar stimuli presented for the second time during the session

Xiang and Brown (1998) found that the neurons with different responses for novel and familiar stimuli can be divided into three categories: novelty neurons that respond strongly only to the first presentations of novel stimuli (N1); recency neurons that respond strongly to stimuli which were not presented recently (N1 and F1); and familiarity neurons that respond strongly to unfamiliar stimuli (N1 and N2). Specific temporal dependencies exist among the activities of the three populations of neurons, suggesting the existence of specific connections between them (Xiang and Brown, 1997; Brown and Xiang, 1998). Bogacz et al. (2001b) have shown that these connections may be efficiently used to increase the reliability of discriminating whether stimuli are being seen for the first time.

The existence of novelty, recency, and familiarity neurons in the perirhinal cortex allows a network to determine not only whether a stimulus is presented for the first time, but also whether it was presented recently (Brown and Xiang, 1998). According to the model proposed by Bogacz et al. (2001b), the neurons of each type create subsystems, each of which could have any of the network architectures for novelty neurons (as discussed earlier, in the section, Description of Networks That Can Perform Familiarity Discrimination). The different behaviours of the neurons in the different subsystems may be reproduced by introducing specialised synaptic properties for recency neurons (synapses that have a shorter memory, e.g., lasting hours, and are reset after a short period of time) and familiarity neurons (synapses that have a delayed or slowly developing plasticity), based on the experimentally observed responses of these neurons (Xiang and Brown, 1998; Brown and Xiang 1998).

The existence of three types of neurons with different temporal characteristics of plasticity may be easily incorporated in the specialised models (Bogacz et al., 2001b). However, their presence seems likely to interfere with the process of feature extraction, assuming the existence of a network combining feature extraction with familiarity discrimination. During feature extraction, a stable code should be learned with neurons representing independent features. Having neurons with delayed plasticity and neurons with short-term plasticity would probably disrupt the stability of the code.

Finally, even within one category (novelty, recency, or familiarity) different neurons have different memory spans (the longest interval after initial presentation of stimuli for which representation of the same stimuli results in a significant change in activity) (Xiang and Brown, 1998). Having neurons with different memory spans inside the network performing feature extraction would probably further disrupt the stability of the code, so providing another argument against the combined models.

In summary, current evidence provides strong, although not yet conclusive, reasons in favour of the existence of a specialised familiarity discrimination network in perirhinal cortex that is separate from any network specialised for feature extraction.

Suggested Experiments

The following experiments establishing the properties of neurons in the perirhinal cortex seem to be the most critical for distinguishing between the models described in this study:

1. To clarify what capacity could be achieved by the human perirhinal cortex working according to various models, it is necessary to find what types of perirhinal neurons have correlated responses. In particular, it would be very useful to know what are the correlations between responses of neurons providing input to the familiarity discrimination network. Answering this question precisely would require measurement of correlations between the responses of the visually responsive neurons that project to the novelty, recency, and familiarity neurons. Besides looking for such correlations amongst groups of simultaneously recorded neurons,

an indirect means of obtaining a closer approximation of these required correlations (than that given by Erickson et al., 2000) is provided by determining the distribution of correlations between perirhinal neurons that are not connected (i.e., which do not have short latency peaks in cross-correlograms, see Gochin et al., 1991). Furthermore, Erickson et al. (2000) found that the distributions of correlation between adjacent perirhinal neurons are different for highly familiar and relatively new stimuli. It would be also interesting to know between which types of perirhinal neurons the changes occur.

2. To provide further support for the anti-Hebbian model, one needs to demonstrate that homosynaptic LTD could be the basis for the decrease of synaptic weights in the anti-Hebbian model. This could be demonstrated by showing that homosynaptic LTD may be induced in some perirhinal neurons by stimulation with a frequency >1 Hz (e.g., by stimulation with patterns such as those recorded from the perirhinal cortex in vivo), or that perirhinal LTD is easier to induce in vivo when animal is exposed to novel visual stimuli (i.e., an experiment analogous to that by Manahan-Vaughan and Braunewell, 1999).

3. To provide further support for the anti-Hebbian model, there is a need to establish whether homosynaptic LTD at some synapses produces an increase in the weights of other synapses of the same neuron, i.e., whether there exists a mechanism of compensatory excitability change that will maintain the excitability of individual perirhinal neurons.

4. To provide further support for the Hebbian inhibitory model, it must be established whether synapses connecting inhibitory and excitatory neurons are plastic.

5. The Hebbian inhibitory model also predicts that inhibitory neurons are stimulus selective. It is known that a large proportion of perirhinal neurons are stimulus selective, but it is not known whether inhibitory neurons are among them. This could be tested after inhibitory neurons have been identified by, for example, matching the shape of action potentials and the firing patterns of perirhinal neurons as has been done for inhibitory and excitatory neurons recorded in the prefrontal cortex (Wilson et al., 1994). Further, identifying the responses of inhibitory neurons in vivo would also allow test of whether some of them are more active for familiar than for novel stimuli, as is predicted by the Hebbian and the combined competitive models.

6. Wang et al. (2000) found that blocking inhibition in inferior temporal cortex causes neurons to start to respond to stimuli to which they did not respond previously, but which produced activity in adjacent neurons. The Hebbian inhibitory model predicts that in a similar experiment (i.e., after blocking inhibition) novelty, recency and familiarity neurons should have the same responses for novel and familiar stimuli, hence such an experiment provides a critical test of the Hebbian inhibitory model.

7. Important evidence would be provided by identifying recency, novelty and familiarity neurons in vitro. If one finds in vitro neurons having synapses with slowly developing plasticity, and other neurons having synapses with rapidly developing plasticity, and the proportion of these neurons matches the one observed in vivo (Xiang and Brown, 1998), then the type and properties of these synapses will distinguish between the anti-Hebbian model, the Hebbian inhibitory model, and the models based on Hebbian learning (the Hebbian model, the combined competitive model, and the double threshold model).

CONCLUSIONS

The calculations and simulations in this study establish that when the responses of neurons providing input to a familiarity discrimination network are correlated (as suggested by experimental data), then specialised networks based on anti-Hebbian learning achieve much larger capacities (up to thousands of times larger) than specialised networks based on Hebbian learning. The currently published combined models do not learn an optimal stimulus representation (they do not fully extract statistically independent features), and hence their capacities are even lower than the specialised models based on Hebbian learning. Hence, the combined models as so far published are critically less efficient than the specialised models based on anti-Hebbian learning.

Currently published experimental observations do not provide sufficient evidence to establish definitively the means by which the perirhinal cortex discriminates familiarity. However, a number of experimental observations favour the existence of a network specialised for familiarity discrimination. Nevertheless, whether familiarity discrimination is performed by a specialised network or by a network that can also perform feature extraction (i.e., one that learns representations of stimuli), this study indicates that the capacity is potentially enormous for the commitment of a very restricted number of neurons. Thus, by providing this small proportion ($<0.1\%$) of the total number of cortical neurons, the brain gains a fast and highly accurate means of detecting things that have not been encountered before and allows neural networks involved in categorisation, association, and recall to be relieved of the need to perform familiarity discrimination. This economy suggests why recognition memory may involve two separable processes, one for familiarity discrimination and one for associative recollection (Bogacz et al., 2001a; Brown and Aggleton, 2001).

Acknowledgments

The authors thank Kenneth Norman and Christophe Giraud-Carrier for discussion and comments on an earlier version of the manuscript, and Cynthia Erickson for providing additional data on the distribution of correlation between responses of distant perirhinal neurons. Calculations in Appendices C and F were done as a student project during an EU Advanced Course in Computational Neuroscience in Trieste (Italy).

REFERENCES

Aggleton JP, Brown MW. 1999. Episodic memory, amnesia and the hippocampal-anterior thalamic axis. Behav Brain Sci 22:425–498.

Aggleton JP, Shaw C. 1996. Amnesia and recognition memory: a re-analysis of psychometric data. Neuropsychologia 34:51–62.
Amit DJ. 1989. Modelling brain function. Cambridge, MA: Cambridge University Press.
Barlow HB. 1989. Unsupervised learning. Neural Comput 1:295–311.
Baxter MG, Murray EA. 2001. Impairment in visual discrimination learning and recognition memory produced by neurotoxic lesions of rhinal cortex in rhesus monkeys. Eur J Neurosci 13:1228–1238.
Bell AJ, Sejnowski TJ. 1997. Edges are the "independent components" of natural scenes. Adv Neural Inform Processing Syst 9:831–837.
Bilkey DK. 1996. Long term potentiation in the in vitro perirhinal cortex displays associative properties. Brain Res 733:297–300.
Bishop CM. 1994. Novelty detection and neural network validation. IEE Proceedings: vision, image and signal processing 141:217–222.
Bliss TVP, Collingridge GL. 1993. A synaptic model of memory: long-term potentiation in hippocampus. Nature 361:31–39.
Bogacz R. 2001. Computational models of familiarity discrimination in the perirhinal cortex. PhD thesis, University of Bristol (available at: http://www.math.princeton.edu/~rbogacz).
Bogacz R, Brown MW, Giraud-Carrier C. 1999. High capacity neural networks for familiarity discrimination. In: Proceedings of international conference on artificial neural networks '99, Edinburgh. p 773–776.
Bogacz R, Brown MW, Giraud-Carrier C. 2000. Frequency-based error back-propagation in a cortical network. In: Proceedings of the International Joint Conference on Neural Networks, Como (Italy). Vol II. p 211–216.
Bogacz R, Brown MW, Giraud-Carrier C. 2001a. Model of familiarity discrimination in the perirhinal cortex. J Comput Neurosci 10:5–23.
Bogacz R, Brown MW, Giraud-Carrier C. 2001b. Model of co-operation between recency, familiarity and novelty neurons in the perirhinal cortex. Neurocomputing 38:1121–1126.
Bogacz R, Brown MW, Giraud-Carrier C. 2001c. Emergence of movement sensitive neurons' properties by learning a sparse code for natural moving images. Adv Neural Inform Processing Syst 13:838–844.
Bogacz R, Brown MW. 2002. The restricted influence of the sparseness of coding on the capacity of the familiarity discrimination networks. Network: Comput Neural Syst 13:457–485.
Brown MW, Aggleton JP. 2001. Recognition memory: what are the roles of the perirhinal cortex and hippocampus? Nat Rev Neurosci 2:51–62.
Brown MW, Xiang JZ. 1998. Recognition memory: neuronal substrates of the judgement of prior occurrence. Prog Neurobiol 55:149–189.
Brown MW, Wilson FAW, Riches IP. 1987. Neuronal evidence that inferotemporal cortex is more important than hippocampus in certain processes underlying recognition memory. Brain Res 409:158–162.
Buckley MJ, Gaffan D. 1998. Perirhinal cortex ablation impairs visual object identification. J Neurosci 18:2268–2275.
Buffalo EA, Reber PJ, Squire LR. 1998. The human perirhinal cortex and recognition memory. Hippocampus 8:330–339.
Burwell RD. 2000. The parahippocampal region: corticocortical connectivity. Ann N Y Acad Sci 911:25–42.
Cho K, Kemp N, Noel J, Aggleton JP, Brown MW, Bashir ZI. 2000. A new form of long-term depression in the perirhinal cortex. Nat Neurosci 3:150–156.
Desai NS, Rutherford LC, Turrigiano GG. 1999. Plasticity in the intrinsic excitability of cortical pyramidal neurons. Nat Neurosci 2:515–520.
Eichenbaum H, Otto T, Cohen NJ. 1994. Two functional components of the hippocampal memory system. Behav Brain Sci 17:449–518.
Erickson CA, Jagadeesh B, Desimone R. 2000. Clustering of perirhinal neurons with similar properties following visual experience in adult monkey. Nat Neurosci 3:1143–1148.
Fahy FL, Riches IP, Brown MW. 1993. Neuronal activity related to visual recognition memory: long term memory and the encoding of recency and familiarity information in the primate anterior and medial inferior temporal and rhinal cortex. Exp Brain Res 96:457–472.
Felleman DJ, Van Essen DC. 1991. Distributed hierarchical processing in the primate cerebral cortex. Cerebr Cortex 1:1–47.
Foldiak P. 1990. Forming sparse representation by local anti-Hebbian learning. Biol Cybern 64:165–170.
Gawne TJ, Richmond BJ. 1993. How independent are the messages carried by adjacent inferior temporal cortical neurons? J Neurosci 13: 2758–2771.
Gerstner W. 1998. Spiking neurons. In: Maas W, Bishop C, editors. Pulsed neural networks. Cambridge, MA: MIT Press. p 4–53.
Gochin PM, Miller EK, Gross CG, Gerstein GL. 1991. Functional interactions among neurons in inferior temporal cortex of the awake macaque. Exp Brain Res 84:505–516.
Grossberg S. 1976. Adaptive pattern classification and universal recoding. I. Parallel development and coding of neural feature detectors. Biol Cybern 23:121–134.
Harpur GF, Prager RW. 1996. Developing of low entropy coding in a recurrent network. Network 7:277–284.
Hertz J, Krogh A, Palmer RG. 1991. Introduction to the theory of neural computations. Redwood City, CA: Addison-Wesley.
Hintzman DL, Caulton DA, Levitin DJ. 1998. Retrieval dynamics in recognition and list discrimination: further evidence of separate processes of familiarity and recall. Mem Cognit 26:449–462.
Holmes P, Lumley JL, Berkooz G. 1996. Turbulence, coherent structures, dynamical systems and symmetry. Cambridge, UK: Cambridge University Press.
Hopfield JJ. 1982. Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci U S A 79:2554–2558.
Insausti R, Juottonen K, Soininen H, Insausti AM, Partanen K, Vainio P, Laakso MP, Pitkanen A. 1998. MR volumetric analysis of the human entorhinal, perirhinal and temporopolar cortices. Am J Neuroradiol 19:659–671.
Ito M. 1989. Long-term depression. Annu Rev Neurosci 12:85–102.
Kemp N, Bashir, ZI. 2001. Long-term depression: a cascade of induction and expression mechanisms. Prog Neurobiol 65:339–365.
Kirov SA, Harris KM. 1999. Dendrites are more spiny on mature hippocampal neurons when synapses are inactivated. Nat Neurosci 2:878–883.
Kirov SA, Sorra KE, Harris KM. 1999. Slices have more synapses than perfusion-fixed hippocampus from both young and mature rats. J Neurosci 19:2876–2886.
Kobatake E, Wang G, Tanaka K. 1998. Effects of shape discrimination training on the selectivity of inferotemporal cells in adult monkeys. J Neurophysiol 80:324–330.
Kohonen T. 1989. Self-organisation and associative memory. 3rd ed. Heidelberg: Springer-Verlag.
Li L, Miller EK, Desimone R. 1993. The representation of stimulus familiarity in anterior inferior temporal cortex, J Neurophysiol 69:1918–1929.
Manahan-Vaughan D, Braunewell KH. 1999. Novelty acquisition is associated with induction of hippocampal long-term depression. Proc Natl Acad Sci U S A 96:8739–8744.
McCulloch WS, Pitts W. 1943. A logical calculus of ideas immanent in nervous activity. Bull Math Biophys 5:115–133.
Meister M, Berry MJ II. 1999. The neural code of the retina. Neuron 22:435–450.
Miller EK, Desimone R. 1994. Parallel neuronal mechanisms for short-term memory. Science 263:520–522.
Miller EK, Li L, Desimone R. 1993. Activity of neurons in anterior inferior temporal cortex during short-term memory task. J Neurosci 13: 1460–1478.
Mitoma H, Konishi S. 1999. Monoaminergic long-term facilitation of GABA-mediated inhibitory transmission at cerebellar synapses. Neuroscience 88:871–883.
Murray EA. 1996. What have ablation studies told us about the neural substrates of stimulus memory? Semin Neurosci 8:13–22.

Murray EA, Bussey TJ. 1999. Perceptual-mnemonic functions of the perirhinal cortex. Trends Cogn Sci 3:142–151.
Norman KA, O'Reilly RC. 2001. Modelling hippocampal and neocortical contributions to recognition memory: a complementary learning systems approach. Technical Report 01-02. Boulder, CO: University of Colorado.
Olshausen BA, Field DJ. 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381: 607–609.
O'Reilly RC, Munakata Y. 2000. Computational explorations in cognitive neuroscience: understanding the mind by simulating the brain. Cambridge MA: MIT Press.
Parra L, Deco G, Miesbach S. 1996. Statistical independence and novelty detection with information preserving non-linear maps. Neural Comput 8:260–269.
Riches IP, Wilson FA, Brown MW. 1991. The effects of visual stimulation and memory on neurons of the hippocampal formation and neighbouring parahippocampal gyrus and inferior temporal cortex of the primate. J Neurosci 11:1763–1779.
Roberts S, Tarassenko L. 1995. A probabilistic resource allocating networks for novelty detection. Neural Comput 6:270–284.
Rolls ET, Baylis GC, Hasselmo ME, Nalva V. 1989. The effect of learning on the face selective responses of neurons in the cortex in the superior temporal sulcus of the monkey. Exp Brain Res 76:153–164.
Seeck M, Michael CM, Mainwaring N, Cosgrove R, Blume H, Ives J, Landis T, Schomer DL. 1997. Evidence for rapid face recognition from human scalp and intracranial electrodes. NeuroReport 8:2749–2754.
Segal M. 2001. Rapid plasticity of dendritic spine: hints to possible functions? Prog Neurobiol 63:61–70.
Sobotka S, Ringo JL. 1993. Investigation of long-term recognition and association memory in unit responses from inferotemporal cortex. Exp Brain Res 96:28–38.
Sobotka S, Ringo JL. 1994. Stimulus specific adaptation in excited but not in inhibited cells in inferotemporal cortex of macaque. Brain Res 646: 94–99.
Sohal VS, Hasselmo ME. 2000. A model for experience-dependent changes in the responses of infero-temporal neurons. Network Comput Neural Syst 11:169–190.
Standing L. 1973. Learning 10,000 pictures. Q J Exp Psychol 25:207–222.
Suzuki WA. 1996. The anatomy, physiology and functions of the perirhinal cortex. Curr Opin Neurobiol 6:179–186.
Suzuki WA, Amaral DG. 1993. Topographic organisation of the reciprocal connections between the monkey entorhinal cortex and the perirhinal and parahippocampal cortices. J Neurosci 14:1856–1877.
Thompson JH, Brown MW, Davies HA, Stewart MG. 2001. Measures of synaptic density in the perirhinal cortex of rats exposed to novel or familiar visual stimuli. Abstract number 537.28. Soc Neurosci Abs 31.
Turrigiano GG, Nelson SB. 2000. Hebb and homeostasis in neuronal plasticity. Curr Opin Neurobiol 10:358–364.
Turrigiano GG, Leslie KR, Desai NS, Rutherford LC, Nelson SB. 1998. Activity-dependent scaling of quantate amplitude in neocortical neurons. Nature 391:892–895.
von der Malsburg C. 1973. Self-organization of orientation sensitive cells in the striate cortex. Kybernetik 14:85–100.
Wang Y, Fujita I, Murayama Y. 2000. Neuronal mechanisms of selectivity for objects features revealed by blocking inhibition in inferotemporal cortex. Nat Neurosci 3:807–813.
Wiesel TN, Hubel DH. 1965. Comparison of the effects of unilateral and bilateral eye closure on cortical unit responses in kittens. J Neurophysiol 28:1029–1040.
Wilson FA, Olscalaidhe SP, Goldman-Rakic PS. 1994. Functional synergism between putative gamma-aminobutyrate-containing neurons and pyramidal neurons in prefrontal cortex. Proc Natl Acad Sci U S A 91:4009–4013.
Xiang JZ, Brown MW. 1997. Processing visual familiarity and recency information: neuronal interactions in area TE and rhinal cortex. Brain Res Assoc Abs 14:69.
Xiang JZ, Brown MW. 1998. Differential neuronal encoding of novelty, familiarity and recency in regions of the anterior temporal lobe. Neuropharmacology 37:657–676.
Xiang JZ, Brown MW. 1999. Differential neuronal responsiveness in primate perirhinal cortex and hippocampal formation during performance of conditional visual discrimination task. Eur J Neurosci 11: 2715–3724.
Ziakopoulos Z, Tillett CW, Brown MW, Bashir ZI. 1999. Input- and layer-dependent synaptic plasticity in the rat perirhinal cortex in vitro. Neuroscience 92:459–472.

APPENDIX A. DESCRIPTION OF THE SIMULATED MODELS

This Appendix describes all the models whose simulations are mentioned in the main text. The two following paragraphs describe the notation used in the Appendix and each of the following subsections describes the details of one of the models.

The notation in the Appendix is similar to that used in previous work on autoassociative memories (Amit, 1989). Assume that a network consists of N novelty neurons, receiving information from N input neurons whose activity pattern represents a visual stimulus. For simplicity assume that each novelty neuron is connected to all the input neurons and denote the strength of the synaptic connection between input neuron j and novelty neuron i by w_{ij}. Denote the activity of input neuron j by x_j, and define the membrane potential of novelty neuron i as:

$$h_i = \sum_{j=1}^{N} w_{ij} x_j \qquad (A.1)$$

Denote the number of presented stimuli (previously stored patterns) by P and the activity of input neuron j after presentation of stimulus μ by x_j^{μ}, so stimulus μ is represented by the pattern of activity of the input neurons given by vector x^{μ}.

A.1. Hebbian Model

In the specialised models analysed and simulated here, the active state of a neuron is denoted by 1, and the inactive state by -1 (such notation is also used in models of associative memories, e.g., Hopfield, 1982). Denoting the inactive state of neurons by -1, rather than 0, simplifies calculation of capacity, and does not change the capacity of familiarity discrimination networks (see Bogacz et al., 2001a). Furthermore, in the specialised models considered here, sparse coding is not assumed, in order to simplify the calculations while still allowing comparisons between the models; i.e., in the following discussion, it is assumed that the probability that an input neuron is active is 50%. Bogacz and Brown (2002) showed that the capacities of familiarity discrimination networks are little influenced by sparseness of coding.

The number of active novelty neurons in familiarity discrimination networks must be limited (see above, Independent Responses

of Novelty Neurons). The Hebbian model assumes for simplicity of calculation that novelty neuron i may be active only if input neuron j is also active. This would correspond to the existence of strong nonmodifiable connections between a novelty neuron and a corresponding input, such that input through this connection is necessary for the novelty neuron to be active. This assumption is made for simplicity of notation and to allow mathematical analysis of the properties of the model. Limiting the number of active novelty neurons may be alternatively achieved by strong connections between groups of neurons rather than between pairs of neurons, or by competition (Bogacz et al., 2001a).

Let d(x) denote the initial network response in the Hebbian model after presentation of input pattern x, i.e., the sum of activities of all novelty neurons just after input reaches the network (this network response determines the level of inhibition in the following period). Let us assume that novelty neurons need to receive input through the strong connections in order to have level of activity larger than zero. Hence, the strong connections ensure that d(x) is equal to the response (proportional to the membrane potential) of neurons receiving input through these strong connections (i.e., $x_i = 1$) (Bogacz et al., 2001a):

$$d(x) = \sum_{i=1}^{N} x_i h'_i = \sum_{\substack{i,j=1 \\ i \neq j}}^{N} x_i w_{ij} x_j \tag{A.2}$$

Thus, d(x) is a dot product of the input pattern and a vector of membrane potentials. The detailed explanation of how such a function may be calculated by a biologically plausible neural network is quite long and hence is not given here but can be found in Bogacz et al. (2001a) and Bogacz and Brown (2002). In Equation A.2, h'_i denotes the membrane potential of novelty neuron i as a result of activity in all connections except the strong connection w_{ii} (the strong connections are assumed to be nonmodifiable, hence their weights do not encode occurrences of the stimuli):

$$h'_i = \sum_{\substack{j=1 \\ j \neq i}}^{N} w_{ij} x_j \tag{A.3}$$

As a result of the Hebbian weight modifications produced by previous occurrences, d is higher for familiar patterns than for novel. In the Hebbian model, d regulates the level of inhibition and hence population activity in the network in the following period (the detailed description of the full network architecture of the Hebbian model is long hence it is not given here but may be found in Bogacz et al., 2001a). Therefore, the familiarity of stimuli may be discriminated reliably by evaluating d or the total network activity in the following period, which is low for high d and high for low d. For simplicity, during simulations described in this study, the biologically plausible network computing d (and activity in the following period) is not simulated explicitly, but instead the familiarity of a stimulus is evaluated by the simulator program computing function d of Equation A.2. The familiarity discrimination threshold may be taken as the middle value between the average decision function d for novel and for familiar stimuli. Before testing the capacity of any network using simulations, the average values of d for novel and familiar stimuli are estimated by presenting to the network 5,000 novel and 5,000 familiar stimuli.

In the Hebbian model, if a presented stimulus x^μ is classified as novel, then in the following period the novelty neurons receiving input via the strong connections fire with high frequency (details of a biologically plausible network resulting in such activity are given in Bogacz et al., 2001a). Hence, in the following, memorising period, the pattern of activity of the novelty neurons is the same as the pattern of activity of input neurons, i.e., it is equal to x^μ. The Hebbian weight modifications after presentation of novel stimulus x^μ may be expressed by modifying every synaptic weight of the novelty neurons by the term:

$$\Delta w_{ij} = \frac{1}{N} x_i^\mu x_j^\mu \tag{A.4}$$

In this rule, synaptic weights are changed according to the activity of the presynaptic input neuron equal to x_j^μ and the postsynaptic novelty neuron equal to x_i^μ: LTP for $x_i^\mu = 1$, $x_j^\mu = 1$; heterosynaptic LTD for $x_i^\mu = 1$, $x_j^\mu = -1$; homosynaptic LTD for $x_i^\mu = -1$, $x_j^\mu = 1$. Equation A.4 also implies that the weights between inactive inputs and inactive novelty neurons ($x_i^\mu = -1$, $x_j^\mu = -1$) should be increased, but Bogacz and Brown (2002) show that it is not a critical element of the model and may be removed. However, the change of weights between inactive neurons is used here to simplify the model, in the same way as for other memory models (Hopfield, 1982; Amit, 1989). In Equation A.4, 1/N is a constant normalising the magnitude of weight modification—as used in associative memories (where the inactive state of a neuron is denoted by -1) (Hopfield, 1982; Amit, 1989).

In simulations, the weights are initialised to 0. The values of weights in the Hebbian model after presentation of P patterns are (Bogacz et al., 2001a):

$$w_{ij} = \frac{1}{N} \sum_{\mu=1}^{P} x_i^\mu x_j^\mu \tag{A.5}$$

According to Equation A.5, the values of the weights of novelty neurons may be either positive or negative; however, it was assumed that the novelty neurons are excitatory (see the section, Description of Networks That Can Perform Familiarity Discrimination). To satisfy the assumption that all novelty neurons are excitatory, all weights may be increased by a constant. In the case in which patterns have different numbers of active bits, the novelty neurons must receive inhibition (additional to any already being used) proportional to the number of active inputs (for details, see Bogacz et al., 2001a). That this modification does not change the capacity is demonstrated by Bogacz et al. (2001a).

A.2. Anti-Hebbian Model

In the anti-Hebbian model (and in the simulated version of the combined competitive model), the number of active novelty neurons is limited not by strong connections but by competition. After presentation of a stimulus μ, the membrane potentials of novelty neurons are calculated according to Equation A.1 and a threshold set such that exactly the half of the novelty neurons with the high-

est membrane potentials are selected to be active. In a real network, such selection of a proportion of the most active neurons may be achieved by inhibition and competition (see, e.g., O'Reilly and Munakata, 2000). The pattern of activity of the novelty neurons after presentation of a stimulus μ is denoted by y^{μ}, namely $y_i^{\mu} = 1$ if neuron i belongs to the group of one-half of the novelty neurons with the highest membrane potential; otherwise $y_i^{\mu} = -1$. The weights of the active novelty neurons are updated according to the rule (illustrated in Fig. 1a):

$$\Delta w_{ij} = -\frac{\eta}{2N}(y_i^{\mu} + 1)x_j^{\mu} \tag{A.6}$$

In Equation A.6, η denotes the learning rate—a parameter determining the magnitude of weight modification; its optimal value depends on N. During simulations, for each size of the network, a learning rate η which gives the highest capacity is found (the capacity given in Fig. 5 is for the optimal η). The optimal value ranged from 0.3 to 0.7 for different networks.

In the simulations of the anti-Hebbian and both combined models, the synaptic weights are initialised to random values. After each weight modification (after simulated stimulus presentations) the weights are normalised such that for each neuron the mean of its weight is 0 and the variance is 1. This normalisation ensures equal chances of activation for each neuron. However, the normalisation also means that stimuli presented initially are not as well remembered as ones presented subsequently. To avoid this problem, the stimuli were presented twice during training in each simulation session, the second presentation being in the reverse order to the first presentation. To investigate the effect of such normalisation and double presentation on capacity, they were also simulated for the Hebbian model (see Bogacz and Brown, 2002): the simulations established that they do not have a large effect on the capacity of the Hebbian model. The double presentation of learning sequences is not essential for the anti-Hebbian model to work. If the weights are normalised as described above and patterns are presented once only, the capacity decreases about two times both for the Hebbian and the anti-Hebbian models, as shown by simulations (not given here). Hence double presentation is introduced here to allow "fair comparison" of capacity with the Hebbian model described in Appendix A.1, in which weights are not normalised after stimulus presentation.

The decision about stimulus familiarity is made by evaluating the following decision function, similar to that of Equation A.2 of the Hebbian model.

$$d(x) = \sum_{i=1}^{N} y_i h_i \tag{A.7}$$

The value of d given by Equation A.7 is larger for novel patterns and lower for familiar, hence the middle value of the average d for novel and for familiar stimuli may be taken as the familiarity criterion.

A.3. Combined competitive model

In the combined models analysed in this study (i.e., the combined competitive and double threshold models), the active state of a neuron is denoted by 1 and the inactive state by 0. This convention is used to allow implementation of the sparse coding necessary for testing feature extraction by the combined models. Let us define the sparseness of coding as the proportion of neurons active in a given pattern or the mean level of activity of the input patterns, and denote it by a, i.e.:

$$a = \frac{1}{N}\sum_{j=1}^{N} x_j^{\mu} \tag{A.8}$$

In the simulated version of the combined competitive model, the number of active novelty neurons is limited by competition (as described for the anti-Hebbian model). After presentation of a stimulus μ, the membrane potentials of novelty neurons are calculated according to Equation A.1 and a threshold set such that precisely the aN novelty neurons with the highest membrane potentials are selected to be active. In a real network, such selection of a number of the most active neurons may be achieved by inhibition and competition (see e.g., O'Reilly and Munakata, 2000). The pattern of activity of the novelty neurons after presentation of a stimulus μ is denoted by y^{μ}. The weights of the active novelty neurons are updated according to the rule (developed from Norman and O'Reilly, 2001):

$$\Delta w_{ij} = \frac{\eta}{Na(1-a)} y_i^{\mu}(x_j^{\mu} - a) \tag{A.9}$$

The expression $1/Na(1-a)$ in Equation A.9 is a simplifying proportionality constant—as used in associative memories (when the inactive state of the neuron is denoted by 0; Amit, 1989).

In the simulations, as for the Hebbian model, the decision about stimulus familiarity was made by evaluating the decision function of Equation A.10 expressing the difference in activity between the most and the least active neurons (for motivation for this decision function see Bogacz and Brown, 2002). As for the Hebbian model, d is larger for familiar than novel patterns, and the value of the decision threshold is taken in simulations as the average of the mean decision function values for novel and familiar stimuli.

$$d(x) = \sum_{i=1}^{N} (y_i - a)h_i \tag{A.10}$$

A.4. Double Threshold Model

In the model as analysed here, the weights are initialised as described at the beginning of section A.2, and after every simulated stimulus presentation, the weights are updated according to the following rule (simplified from Sohal and Hasselmo, 2000; the original equation for the weight updating rule in their study is very complex; the simplifications made below do not change the operational principles of the model and hence its capacity, while making possible its mathematical analysis):

$$\Delta w_{ij} = \frac{\eta}{Na(1-a)}(y_i^{\mu} - a)(x_j^{\mu} - a) \tag{A.11}$$

In Equation A.11, y_i^{μ} is equal to 1 if neuron i is above a plasticity threshold, and is 0 otherwise. In the original Sohal and Hasselmo

(2000) model the plasticity threshold is fixed. But here for simplicity of analysis, assume that exactly aN novelty neurons with the highest membrane potentials are above the plasticity threshold. This would mean that the plasticity threshold is chosen for each stimulus such that the above criterion is satisfied. Although this assumption is not easily implemented biologically, it simplifies the calculations and, importantly, in this model it does not decrease the capacity of the network (see Bogacz and Brown, 2002).

During simulations, the decision about stimulus familiarity is made by evaluating a decision function that is equal to the number of the novelty neurons with membrane potential above the activation threshold of $-a/2$ (see Bogacz and Brown, 2002). As for previously described models, the value of the decision threshold is taken in simulations as the average of the mean decision function values for novel and familiar stimuli.

APPENDIX B. FEATURE EXTRACTION BY COMBINED MODELS

Appendix B.1. Proportion of Missing Features

This Appendix calculates the average fraction of features that will be missed by a feature extraction network in which neurons choose independently of one another features to represent, as in the combined competitive and double threshold networks.

Let us denote the number of independent features from which the input patterns are built by F, and the number of the output neurons (that learn to represent features) by M. Let us assume that each neuron has learnt to represent one feature and this feature was selected independently from those selected by other neurons. The probability that neuron i represents feature j is $1/F$; hence the probability that feature j is not represented by neuron i is equal to $1-1/F$. Therefore, the probability that feature j is not represented by any neuron is $(1-1/F)^M$, so the average number of missed features is $M(1-1/F)^M$, and the average fraction of missed features is $(1-1/F)^M$.

Let us assume that there are k times more neurons than features, i.e., $M = kF$; thus, the average fraction of missed features is $(1-1/F)^{kF}$. Let us calculate the limit of this fraction as F goes to infinity:

$$\lim_{F\to\infty}\left(1-\frac{1}{F}\right)^{kF} = \left[\lim_{F\to\infty}\left(\frac{F}{F-1}\right)^{F}\right]^{-k} \overset{n=F-1}{=} \left[\lim_{n\to\infty}\frac{n+1}{n}\left(\frac{n+1}{n}\right)^{n}\right]^{-k} = e^{-k} \quad \text{(B.1)}$$

In Equation B.1, the transformation from the first to the second line involves a change of variables. The term $(n+1)/n$ in the expression in the second line converges to 1, so it may be discarded. The limit is then equal to the definition of the Euler constant e. Since the sequence defining e converges very rapidly, e^{-k} is a very precise approximation of $(1-1/F)^{kF}$ for larger F (e.g., $F > 100$).

To summarise, this Appendix shows that for a feature extraction network in which neurons choose features to represent independently from one another, and which has k times more neurons than independent features, underlying input patterns will on average miss a proportion of approximately e^{-k} features.

B.2. Patterns Used for Feature Extraction

Patterns used for testing the ability of combined models to extract features (see the section, Feature Extraction by Combined Models) were generated in the following way. At the beginning of the simulation session, M independent features F^m were generated; these were random binary sequences with 5 bits equal to 1 and the rest equal to 0 (M was also equal to the number of novelty neurons, and in the simulations $M = N$). To ensure that an equal probability of different input neurons is active, the following constraint was forced:

$$\forall\, i \sum_{m=1}^{M} F_i^m = const \quad \text{(B.2)}$$

In the case in which all the features were equally strong (i.e., strength = 1), each pattern x^μ was a sum of five randomly chosen F^m, so x_j^μ were natural rather than binary numbers. Let us denote the binary vector describing the independent features present in pattern x^μ by Y^μ (thus, in Y^μ there are 5 values equal to 1 and the rest of the values are equal to 0) so that:

$$x_j^\mu = \sum_{m=1}^{M} Y_m^\mu F_j^m \quad \text{(B.3)}$$

In the case in which some features were stronger than others, in the feature vectors F^m of the 10 strong features ($m \in \{1,\ldots 10\}$) all the 1s were replaced by the value of the parameter strength describing the strength of the strong features relative to the others. As for the case of equally strong features, the constraint of Equation B.2 was forced to ensure an equal probability that different input neurons would be active, and the patterns were generated according to Equation B.3.

Afterwards, to ensure appropriate sparseness of coding, each pattern was normalised such that the constraint of Equation A.8 was forced (after this normalisation x_j^μ were real numbers). The sparseness of coding was taken as $a = 0.1$, because 5 of 50 features were present in each pattern (i.e., 10%) hence 10% of novelty neurons should be active when the optimal representation is learned (i.e., when each novelty neuron represents one feature).

Before matching the weights of novelty neurons to the features, the features were normalised in the same way as the weights of the novelty neurons in the combined competitive model (see the beginning of section A.2), such that for each feature m, the mean of F_j^m is 0, and the variance is 1. Then, for each neuron i, the Euclid distances were compared between the weight vector w_i and all the normalised feature vectors F^m, and it was considered that the neuron i represented the feature F^m that had the shortest Euclid distance to w_i.

TABLE 1.

Elements of Summation in Equation C.2

No.	Case	No. of cases	Average of elements
1	$i = k \wedge j = l \wedge \mu = \theta$, or $i = l \wedge j = k \wedge \mu = \theta$	$2N^2P$	1
2	$i = k \wedge j = l \wedge \mu \neq \theta$, or $i = l \wedge j = k \wedge \mu \neq \theta$	$2N^2P$	r_{ij}^2
3	$(i = k \vee j = 1 \vee i = l \vee j = k)$, and $\mu = \theta$	$4N^3P$	r_{ij}^2 (for $i = k$)
4	$(i = k \vee j = l \vee i = l \vee j = k)$, and $\mu \neq \theta$	$4N^3P^2$	$r_{ij}r_{kl}r_{jl}$ (for $i = k$)
5	Otherwise	The rest	$r_{ij}^2r_{kl}^2$

APPENDIX C. STORAGE CAPACITY OF THE HEBBIAN MODEL FOR CORRELATED PATTERNS

C.1. Derivation of Capacity

In this Appendix, we calculate the capacity of the Hebbian model using signal-to-noise analysis (for a clear introduction to this method see Hertz et al., 1991), as follows. First, we calculate: (1) the mean value of the decision function d (defined in Equation A.2) for familiar patterns, (2) the mean d for novel patterns, and (3) the variance of d across patterns. Using these three values, we find the probability of discrimination error. Finally, having the expression for error probability, we find the number of stored patterns P_{max} for which this error probability is 1%; P_{max} is then the capacity.

We assume that the patterns reflect some regularities in the external world and hence the activities of some input neurons are correlated. But for simplicity we assume that the average value of each bit of a pattern is 0, i.e., $\langle x_j^\mu \rangle_\mu = 0$. Hence the correlation between inputs i and j is equal to $r_{ij} = \langle x_i^\mu x_j^\mu \rangle_\mu$. Denote the matrix of r_{ij} by **R** (for random uncorrelated patterns, **R** is the identity matrix).

Let us calculate the value of the decision function (defined in Equation A.2) after presentation of the first familiar stimulus.

$$d(x^1) = \sum_{i=1}^{N} x_i^1 h_i = \sum_{\substack{i,j=1 \\ i \neq j}} x_i^1 x_j^1 w_{ij} = \frac{1}{N} \sum_{\substack{i,j=1 \\ i \neq j}}^{N} x_i^1 x_j^1 x_i^1 x_j^1 + \frac{1}{N} \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \sum_{\mu=2}^{P} x_i^1 x_j^1 x_i^\mu x_j^\mu \quad \text{(C.1)}$$

The first term in the last line of Equation C.1 is called "signal" in neural network literature, and the second—"noise" (Amit, 1989). Since $x_j^1 \in \{-1,1\}$, $x_j^1 x_j^1 = 1$, and the signal term is equal to N. The expected values of the expressions $x_i^1 x_j^1 x_i^\mu x_j^\mu$ in the noise term are equal to r_{ij}^2. Let us denote the average of the squares of the correlations between the neurons' activities by $r^2 = \langle r_{ij}^2 \rangle_{i \neq j}$ (for biased patterns used in the section, Capacity for Correlated Input Patterns, $r^2 = b^4$). Hence, the average value of the noise is equal to NPr^2 (P denotes the number of stored patterns). To calculate the variance of the noise, let us find the average of the square of the noise term:

$$Noise^2 = \frac{1}{N^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \sum_{\mu=2}^{P} \sum_{\substack{k,l=1 \\ k \neq l}}^{N} \sum_{\theta=2}^{P} x_i^1 x_j^1 x_i^\mu x_j^\mu x_k^1 x_l^1 x_k^\theta x_l^\theta \quad \text{(C.2)}$$

Table 1 shows all the possible values of the elements of the summation of Equation C.2, depending on equalities between indices. Cases 2 and 3 (Table 1) contribute little to the variance of the noise (because they occur much less frequently than cases 4 and 5), and discarding them does not change the estimation of capacity materially (Fig. 5a). If only cases 1, 4, and 5 are considered, then the average of the noise squared is equal to $2P+4NP^2r^3+(NPr^2)^2$, where r^3 is defined by Equation C.3 (for biased patterns used in the section, Capacity for Correlated Input Patterns, $r^3 = b^6$).

$$r^3 = \langle r_{ij} r_{il} r_{jl} \rangle_{i \neq j \neq l} \quad \text{(C.3)}$$

Hence, given the above paragraph, the variance of the noise is equal to $2P+4NP^2r^3$ (because $D^2(\text{Noise}) = \langle \text{Noise}^2 \rangle - \langle \text{Noise} \rangle^2$).

The above analysis indicates that the decision function after presentation of a familiar stimulus has mean $N+NPr^2$ and variance $2P+4NP^2r^3$, while after presentation of a novel stimulus the mean is NPr^2 (because there is no signal) with the same variance. Hence $d > N/2+NPr^2$ may be used as a familiarity criterion (i.e., a population activity of novelty neurons as measured by the decision function of Equation A.2 above $N/2+NPr^2$ would indicate that the stimulus is novel, an activity below it, that the stimulus is familiar). Since it was assumed that all the patterns have the same regularities, r^2 does not change in time and the familiarity discrimination threshold may be set during brain development. We consider the network as working well if the probability of error is less than 1%. An error occurs if the noise is higher than the threshold. To calculate the maximum acceptable number of stored patterns P_{max}, we must solve the following equation:

$$\Pr\left[\theta(NP_{max}r^2,\ \sqrt{2P_{max} + 4NP_{max}^2r^3}) < NP_{max}r^2 + \frac{N}{2}\right] = 0.99 \quad \text{(C.4)}$$

In Equation C.4, Pr denotes probability. Equation C.4 is equivalent to:

$$\Pr\left[\theta(0,\ 1) < \frac{N}{\sqrt{8P_{max} + 16NP_{max}^2r^3}}\right] = 0.99 \quad \text{(C.5)}$$

Since the noise may be estimated by a normal distribution, Equation C.5 may be solved by checking the value of the inverted standard normal cumulative distribution for 0.99:

$$\frac{N}{\sqrt{8P_{max} + 16NP_{max}^2 r^3}} \approx 2.33 \quad \text{(C.6)}$$

Solving Equation C.6 with respect to P_{max}, we get

$$P_{max} = \frac{-1 + \sqrt{1 + 0.185N^3 r^3}}{4Nr^3} \quad \text{(C.7)}$$

Equation C.7 shows that the presence of correlations both between inputs and between novelty neurons reduces capacity very markedly: even for small values of r^3, for very large N, P_{max} is proportional to $\sqrt{N}$, rather than N^2.

Equation C.7 gives the capacity for the simplified case when the network is fully connected. Let us calculate the capacity when the connections between perirhinal neurons are sparse (i.e., each novelty neuron receives inputs only from a small proportion of input neurons), which is more likely to reflect the real situation in the perirhinal cortex. Let us denote the probability of an input neuron being connected to a novelty neuron by c. Hence, on average, Nc weights of a novelty neuron have the same values as defined in Equation A.5, and the rest of the weights are equal to 0 (which corresponds to a lack of connection). For simplicity, we still do not consider any spatial organisation of the connections; i.e., let us assume that the probability of a novelty neuron receiving a connection from an input neuron is equal for all inputs (no matter what the distance between the input neuron and the novelty neuron). Although this is a simplification, it is at least partially consistent with the pattern of intrinsic connections in the perirhinal cortex (see above, Combined or Specialised Network?).

Let us investigate how the signal and noise in the decision function change due to such sparse connectivity. Calculations analogous to those of Equation C.1 show that the average value of the signal becomes equal to Nc. To calculate the variance of the noise it is necessary to analyse the values of the elements of summation from Table 1. In cases in which $i = k \wedge j = l$, the average values of the elements should be multiplied by c, and in all other cases the average values of the elements should be multiplied by c^2. Hence, if, as previously, only cases 1, 4, and 5 from Table 1 are taken into consideration, then the variance of the noise becomes $Pc + Pc^2 + 4NP^2c^2r^3$. Calculations of capacity analogous to those of the previous part of this Appendix show that the capacity is equal to:

$$P_{max} = \frac{-1 - c + \sqrt{(1 + c)^2 + 0.74N^3c^2r^3}}{8Ncr^3} \quad \text{(C.8)}$$

Equation C.8 shows that even when r^3 is quite small, P_{max} is proportional to $\sqrt{N}$, rather than to the number of synapses N^2c. A sparse connectivity reduces the impact on capacity of correlation within patterns: although Equation C.8 is complex and hence difficult to interpret directly, the reduced impact of correlation as sparseness of connections increases is clearly visible in Figure 5b.

Associative memories that have ability to recall information also show a decrease in capacity when patterns to be stored are correlated (Hertz et al., 1991). In the case of associative memories, the influence of correlation on capacity may be reduced by repeating presentations of the stimuli (or the pseudo-inverse rule; Hertz et al., 1991). However, this approach does not work for the familiarity discrimination networks. Ignoring the fact that what is wanted is single exposure learning, Bogacz et al. (2000) have shown that repeating presentations of stimuli to a familiarity discrimination network, indeed reduces the probability of classifying a familiar stimulus as novel but does not reduce the probability of error of classifying a novel stimulus as familiar. Hence the overall probability of a familiarity discrimination error, which is the average of the two above probabilities (when it is assumed that the novel and familiar stimuli occur equally often) may be reduced by no more than two times by repeating stimulus presentations. Hence repeating stimulus presentations may increase the capacity slightly (Bogacz et al., 2000), but not sufficiently to overcome the decrease in capacity due to correlation in the input patterns.

To summarise, this Appendix establishes that any correlation between the responses of the input neurons reduces the capacity of the Hebbian model very markedly: even for relatively small values of correlation, the capacity becomes proportional to $\sqrt{N}$, rather than to the number of synapses. Correspondingly, in any real network based on such computational principles, it is essential for correlation between neuronal responses to be very low, if capacity is to be maximised.

C.2. Estimation of Correlation Between Responses of Perirhinal Neurons

Appendix C.1 calculates how the capacity of the Hebbian model depends on the correlations between input neurons. However, the capacity that may be achieved by the Hebbian model remains undetermined until parameter r^3 is known. The value of r^3 likely for activities of real perirhinal neurons is estimated in this section.

Using Monte Carlo methods, it is possible to estimate the most likely real underlying distribution of correlations between distant perirhinal neurons, i.e., its mean $r\mu$ and standard deviation $r\sigma$. Simulations with different values of $r\mu$ and $r\sigma$ were used to determine the distribution of correlations between responses of distant perirhinal neurons most likely to result in the distribution of estimated correlations observed by Erickson et al. (2000) (see above, Correlation Between Responses of Real Perirhinal Neurons). In each simulation session, we calculated the estimated correlation between 10^7 pairs of simulated neurons. For each pair we took the correlation r as a random number taken from the normal distribution with mean $r\mu$ and standard deviation $r\sigma$. Then, we generated 16 pairs of random numbers (corresponding to the pairs of responses to the 16 stimuli used in Erickson et al.'s, experiment) taken from two distributions with correlation r, and we estimated correlation $\hat{r}$ between these 16 pairs of numbers. For each session, we found the mean $\hat{r}\mu$ and standard deviation $\hat{r}\sigma$ of $\hat{r}$. These numerical simulations showed that the most likely distribution of estimated correlations with mean $\hat{r}\mu \approx 0.05$ and standard deviation $\hat{r}\sigma \approx 0.313$ (the values observed by Erickson et al., 2000), is obtained when the random numbers are generated according to random distributions with correlations with mean $r\mu \approx 0.05$ and standard deviation $r\sigma \approx 0.21$.

In the above calculations, the correlation r was estimated from 16 pairs of numbers, rather than 24 (in the experiment carried out

by Erickson et al., correlations were calculated between mean neuronal responses to presentations of 16 or 24 stimuli) to produce a more conservative estimate of correlation distribution (i.e., lower values of $r\mu$ and $r\sigma$).

To estimate the value of r^3 based on the above calculation, we generated 10^7 triplets of random numbers drawn from a normal distribution with mean $r\mu \approx 0.05$ and standard deviation $r\sigma \approx 0.21$ and estimated r^3 using definition of Equation C.3, and obtained $r^3 \approx 0.0001646$. The cube root of the estimated r^3 is equal to 0.055, so it is close to $\hat{r}\mu \approx 0.05$; the latter value is used in estimation of Figure 7.

APPENDIX D. UPPER LIMIT OF CAPACITY OF COMBINED MODELS

Appendix C calculated the capacity of the Hebbian model for correlated input patterns. Since it was assumed for simplicity of calculation that the pattern of activity of the novelty neurons was the same as the pattern of activity of the input neurons (due to one-to-one driving connections), the correlations between activities of the novelty neurons were the same as between the inputs. This Appendix calculates (in an analogous way) the upper bound of the capacity of the combined models, with the assumption that they perform perfect feature extraction, i.e., the activities of the novelty neurons are statistically independent (hence uncorrelated). Therefore, the pattern of activity of the novelty neurons after presentation of stimulus μ must be different from the pattern of activity of the input neurons, because the novelty neurons' activities are not correlated (while the input neurons' are). Let us denote the pattern of activity of the novelty neurons after presentation of stimulus μ by y^μ and the pattern of activity of the input neurons by x^μ (as previously).

As in Appendix C, let us define the correlation between inputs i and j as $r_{ij} = \langle x_i^\mu x_j^\mu \rangle_\mu$ and denote the matrix of r_{ij} by $\mathbf{R}$. In addition, let us define the correlation between novelty neurons i and j as $\dot{r}_{ij} = \langle y_i^\mu y_j^\mu \rangle_\mu$, the correlation between novelty neuron i and input neuron j as $\hat{r}_{ij} = \langle y_i^\mu y_j^\mu \rangle_\mu$, and denote the matrices of $\dot{r}_{ij}$ and $\hat{r}_{ij}$ by $\dot{\mathbf{R}}$ and $\hat{\mathbf{R}}$, respectively. In Appendix C, it was assumed that all these matrices were equal: $\mathbf{R} = \dot{\mathbf{R}} = \hat{\mathbf{R}}$. By contrast, here we assume that $\dot{\mathbf{R}}$ is the identity matrix, which means that the activities of the novelty neurons are not correlated (the correlation between adjacent neurons is not considered here).

For simplicity, let us first consider the case of a fully connected network (i.e., each novelty neuron receives connections from each input neuron). The values of the weights of the novelty neurons after presentation of P stimuli become:

$$w_{ij} = \frac{1}{N} \sum_{\mu=1}^{P} y_i^\mu x_j^\mu \tag{D.1}$$

Term y_i^μ replaces x_i^μ also in the decision function. Let us calculate the decision function after presentation of the first familiar stimulus:

$$d(x^1) = \sum_{i=1}^{N} y_i^1 h_i = \sum_{i,j=1}^{N} y_i^1 x_j^1 w_{ij} =$$

$$\frac{1}{N} \sum_{i,j=1}^{N} y_i^1 x_j^1 y_i^1 x_j^1 + \frac{1}{N} \sum_{i,j=1}^{N} \sum_{\mu=2}^{P} y_i^1 x_j^1 y_i^\mu x_j^\mu \tag{D.2}$$

As in Appendix C, the signal is equal to N. The average value of expression $y_i^1 x_j^1 y_i^\mu x_j^\mu$ in the noise term is $\hat{r}_{ij}^2$; hence the average value of the noise is equal to $NP\langle \hat{r}_{ij}^2 \rangle_{i,j}$. To calculate the variance of the noise, we analyse the square of the noise:

$$Noise^2 = \frac{1}{N^2} \sum_{i,j=1}^{N} \sum_{\mu=2}^{P} \sum_{k,l=1}^{N} \sum_{\theta=2}^{P} y_i^1 x_j^1 y_i^\mu x_j^\mu y_k^1 x_l^1 y_k^\theta x_l^\theta \tag{D.3}$$

The averages of the elements of the summation in the noise term have different values for different configurations of indices (parallel to the analysis of Table 1 for the Hebbian model). However, since we are calculating the upper limit of capacity let us consider only those cases which contribute most to the variance of the noise. Taking into consideration other cases as well would result in a slightly bigger value of the noise and hence in a slightly lower capacity (therefore, the calculation will slightly overestimate the upper limit of capacity). Let us consider the three cases corresponding to cases 1, 4, and 5 from Table 1 that were considered during the calculation of capacity in Appendix C. First, when $i = k \wedge j = l \wedge \mu = \theta$ (there are N^2P such cases), all the elements are equal to 1. This case corresponds to the noise which is present independent of correlation. Second, when $i = k \wedge \mu \neq \theta$ (there are N^3P^2 such cases), the average of the elements is $r_{jl}\hat{r}_{ij}\hat{r}_{il}$. This expression is different from zero when there exist the following correlations: between inputs j and l, between input j and novelty neuron i, and between input l and novelty neuron i. Note that if the network is doing feature extraction, such correlations must exist for some triplets of neurons. In particular, if the feature encoded by novelty neuron i is represented by the activities of input neurons j and l, the activities of inputs j and l, and novelty neuron i are correlated because they encode the same feature. Thus, whenever the feature is present, all three neurons are coactive. Third, when no equality between indices is present (corresponding to case 5 of Table 1), the average of the element of summation in Equation D.3 is $\hat{r}_{ij}^2\hat{r}_{kl}^2$. Considering these three cases, it may be shown that the variance of the noise is $P+NP^2\hat{r}^3$ where $\hat{r}^3 = \langle r_{ij}\hat{r}_{ij}\hat{r}_{il} \rangle_{ij,l}$. Knowing this, the capacity may be calculated using the technique from the previous Appendix, giving

$$P_{\max} \approx \frac{-1 + \sqrt{1 + 0.185N^3\hat{r}^3}}{2N\hat{r}^3} \tag{D.4}$$

Similar calculations for the case of sparse connections give:

$$P_{\max} \approx \frac{-1 + \sqrt{1 + 0.185N^3c^2\hat{r}^3}}{2Nc\hat{r}^3} \tag{D.5}$$

Equations D.4 and D.5 are visually similar to Equations C.7 and C.8. They also establish that even for relatively small $\hat{r}^3$,the capac-

ity of combined models is proportional to $\sqrt{N}$, rather than to the number of synapses. However, note that the upper bounds of capacity of the combined models (Equations D.4 and D.5) are higher than the capacities of the Hebbian model (Equations C.7 and C.8) for the same correlation in input patterns (expressed by matrix **R**). This follows because $\hat{r}^3 < r^3$ since it was assumed that the responses of novelty neurons were not correlated. However, when the responses of novelty neurons are correlated (i.e., the network does not complete feature extraction), then $\hat{r}^3$ increases; furthermore, the elements of summation of Equation D.3 will also have large values for some other combinations of indices (in particular $j = l \wedge \mu \neq \theta$), which decreases the capacity.

Establishing the relative capacities of the Hebbian and the upper limit of the combined models requires an understanding of the relationship between $\hat{r}^3$ and r^3. No precise analytical solution determining this relationship for all possible feature extraction networks has been found. However, a very approximate relationship between $\hat{r}^3$ and r^3 has been found in Bogacz (2001). Analysing approximations shows two things about $\hat{r}^3$. First, $\hat{r}^3$ is smaller than r^3 (Bogacz, 2001). Hence, comparing Equations D.4 and C.7, it is clear that for correlated input patterns, networks which complete feature extraction achieve a larger capacity than networks that do not complete feature extraction. Second, if r^3 grows, $\hat{r}^3$ also grows (proportionally) (Bogacz, 2001). Hence, a specialised network receiving uncorrelated inputs will achieve a higher capacity than a network of the same size receiving correlated inputs, even if the latter completes feature extraction. Bogacz (2001) describes also an attempt to find the upper limit of capacity of the combined models in simulations. The above two predictions are consistent with the simulations in Bogacz (2001).

The upper limit of capacity calculated in this Appendix may not be the upper limit of capacity of all possible familiarity discrimination networks. However, it is the upper limit for the class of single layer networks based on Hebbian learning because the responses of the novelty neurons from the model analysed in this Appendix are uncorrelated, and the capacity will decrease if the novelty neurons become correlated. Any combined model that achieved a larger capacity would have to operate on different principles from those investigated here. At present the development of such a model with biologically plausible learning rules seems improbable.

To summarise, this Appendix calculates the upper bound of capacity for the class of combined models that are single layer networks with Hebbian learning. It shows that when there are even small correlations in the input patterns, the capacity of such combined models is proportional to $\sqrt{N}$, rather than to the number of synapses. However, if a combined network completes feature extraction, the capacity of the combined model can be larger than the capacity of the specialised Hebbian model for input patterns having the same correlation. Nevertheless, when the activities of novelty neurons become correlated (i.e., the network fails to complete feature extraction), the capacity of the combined models decreases and so falls below the upper limit calculated in this Appendix.

APPENDIX E. STORAGE CAPACITY OF THE ANTI-HEBBIAN MODEL FOR CORRELATED PATTERNS

The behaviour of the anti-Hebbian model is very complex, hence its capacity is calculated in this Appendix using the following approximations.

1. During calculations, the values of various constants are discarded, and thus the Appendix finds analytically only the qualitative relation between the capacity P and the size of the network N, rather than an exact equation for capacity. Hence we use sign "$\approx$" to denote that two values are approximately equal when the precise value of a constant has not been determined or that no more than an approximate proportionality has been established.
2. For simplicity, only a fully connected network is considered here.
3. The calculations are done only for patterns biased towards the template (see those used earlier, as described in the section, Capacity for Correlated Input Patterns). A general expression for capacity (i.e., valid for any patterns) has not been found, and finding it has proved very difficult. In addition, we assume that the template is not switched at random moments in time (see above, Capacity for Correlated Input Patterns). This difference in the method of pattern generation does not have an influence on capacity of the anti-Hebbian model, as shown by the results of simulations.
4. This Appendix calculates the capacity of a modification of the anti-Hebbian model, in which the weights of inactive neurons are also modified. That is, in the network analysed here the weights are modified according to the following equation instead of Equation A.6.

$$\Delta w_{ij} = -\frac{\eta}{N} y_i^\mu x_j^\mu \tag{E.1}$$

The learning rule of Equation E.1 would be difficult to implement in a biological neural network (see Bogacz and Brown, 2002), but this modification of the learning rule from that of Equation A.6 simplifies the calculation of capacity, and does not change the capacity significantly from that of the more realistic rule (as will be shown in simulations described in this Appendix).

The weights in the anti-Hebbian model after presentation of P patterns are approximately proportional to

$$w_{ij} \approx -\frac{1}{N} \sum_{\mu=1}^{P} y_i^\mu x_j^\mu \tag{E.2}$$

In Equation E.2, there is an approximate equality, because we have discarded (1) the fact that the weights were initialised to random values, not to 0; (2) the constant η; and (3) the weight normalisation process described in Appendix A.2.

To find the capacity of the anti-Hebbian model, we will use the signal-to-noise analysis as in Appendices C and D. Let us calculate the value of the decision function (defined in Equation A.7) after presentation of the first stimulus:

$$d(x^1) \approx \sum_{i=1}^{N} y_i^1 h_i = \sum_{i,j=1}^{N} y_i^1 x_j^1 w_{ij} \approx$$

$$= -\frac{1}{N} \sum_{i,j=1}^{N} y_i^1 x_j^1 y_i^1 x_j^1 - \frac{1}{N} \sum_{i,j=1}^{N} \sum_{\mu=2}^{P} y_i^1 x_j^1 y_i^\mu x_j^\mu \quad \text{(E.3)}$$

In the first line of Equation E.3, there is an approximate equality after d, because it is assumed that there is the same pattern y^1 during two different presentations of the first stimulus; i.e., it is assumed that the same novelty neurons belong to the group of the most active neurons (or more precisely the same neurons have their membrane potentials among N/2 highest membrane potentials in the network). However, in the anti-Hebbian model, patterns y^1 during two different presentations of a stimulus usually only partially overlap.

Calculations analogous to those of Equation C.1 show that the signal term is equal to $-N$. To calculate the distribution of the noise term, rewrite it as

$$Noise = -\sum_{i=1}^{N} y_i^1 \sum_{\mu=2}^{P} y_i^\mu \frac{1}{N} \sum_{j=1}^{N} x_j^1 x_j^\mu \quad \text{(E.4)}$$

We may then calculate the distribution of the noise term from Equation E.4, starting with the terms in the innermost summation.

Since the patterns are biased towards the same template and $\Pr(x^\mu = x^{temp}) = \frac{1}{2}(1+b)$ (see above, Capacity for Correlated Input Patterns), then

$$\langle x_j^1 x_j^\mu \rangle = \frac{(1+b)^2}{4} + \frac{(1-b)^2}{4} - 2\frac{(1+b)(1-b)}{4} = b^2 \quad \text{(E.5)}$$

Since $x_j^\mu \in \{-1,1\}$, $(x_j^1 x_j^\mu)^2$ is equal to 1. Hence the term $x_j^1 x_j^\mu$ has variance equal to $1-b^4 \approx 1$. Thus, the innermost summation of Equation E.4 may be approximated by the following normal distribution:

$$\frac{1}{N} \sum_{j=1}^{N} x_j^1 x_j^\mu \approx \theta\left(b^2, \sqrt{\frac{1}{N}}\right) \quad \text{(E.6)}$$

Let us now estimate the distribution of term

$$\sum_{\mu=2}^{P} y_i^\mu$$

From Equation E.4. Note that this term describes how many times neuron i was activated during presentation of P stimuli. In particular, if the neuron was activated for exactly one-half of the stimuli, the term is equal to 0; if the neuron was activated for more than one-half of the stimuli, the term is positive; and if the neuron was activated for less than one-half of the stimuli, the term is negative. Let us define a term s_i^q describing how many times neuron i was activated during presentation of first q stimuli:

$$s_i^q = \sum_{\mu=1}^{q} y_i^\mu \quad \text{(E.7)}$$

Let us investigate how s_i^{q+1} depends on s_i^q and q. From Equation E.7 we get

$$s_i^{q+1} = s_i^q + y_i^{q+1} \quad \text{(E.8)}$$

Hence, let us calculate how the probability distribution of y_i^{q+1} depends on s_i^q and q. Let us recall that in the anti-Hebbian model, y_i^{q+1} is equal to 1 when novelty neuron i has membrane potential $h_i(x^{q+1})$, which is among N/2 highest membrane potentials in the network; and y_i^{q+1} is equal to -1 otherwise. From Equation A.1, we can expect that the membrane potentials of novelty neurons are normally distributed with mean 0, hence for large N, for the great majority of novelty neurons:

$$y_i^{q+1} = \mathrm{sgn}[h_i(x^{q+1})] \quad \text{(E.9)}$$

Hence, let us calculate how the probability distribution of $h_i(x^{q+1})$ depends on s_i^q and q:

$$h_i(x^{q+1}) = \sum_{j=1}^{N} w_{ij} x_j^{q+1} \approx -\sum_{\mu=1}^{q} y_i^\mu \frac{1}{N} \sum_{j=1}^{N} x_j^\mu x_j^{q+1} \quad \text{(E.10)}$$

From Equation E.6, it follows that Equation E.10 is equivalent to

$$h_i(x^{q+1}) \approx -\sum_{\mu=1}^{q} y_i^\mu \theta\left(b^2, \sqrt{\frac{1}{N}}\right) \quad \text{(E.11)}$$

From the definition of Equation E.7, it follows that Equation E.11 is equivalent to:

$$h_i(x^{q+1}) \approx \theta\left(-s_i^q b^2, \sqrt{\frac{q}{N}}\right) \quad \text{(E.12)}$$

From Equations E.9 and E.12, we get

$$\Pr(y_i^{q+1} = 1) = \Pr[h_i(x^{q+1}) > 0] \approx \Pr\left[\theta\left(-s_i^q b^2, \sqrt{\frac{q}{N}}\right) > 0\right]$$

$$= \Pr\left[\theta(0,1) < -s_i^q b^2 \sqrt{\frac{N}{q}}\right] = \Phi\left(-s_i^q b^2 \sqrt{\frac{N}{q}}\right) \quad \text{(E.13)}$$

In Equation E.13, Φ denotes the cumulative standard normal distribution. Using Taylor's expansion in vicinity of 0 we get

$$\Phi(x) \approx \frac{1}{2} + \frac{1}{\sqrt{2\pi}} x$$

Hence we can approximate

$$\Pr(y_i^{q+1} = 1) \approx \frac{1}{2} - s_i^q \left(b^2 \sqrt{\frac{N}{2\pi q}}\right) \quad \text{(E.14)}$$

The interpretation of Equation E.14 is that if $s_i^q > 0$ (i.e., the neuron was active for more than one-half the presented stimuli), the probability that the neuron will respond to the next stimulus is lower than $\frac{1}{2}$. Furthermore, the larger s_i^q is (i.e., the more the neuron was active in the past), the lower the probability that the neuron

will respond in the future. Conversely, if $s_i^q < 0$ (i.e., the neuron was active for less than one-half the presented stimuli), the probability of neuron to respond again is $> \frac{1}{2}$. Hence the novelty neurons in the anti-Hebbian model have a natural tendency to be active for exactly one-half of the stimuli.

Intuitively, the reason for this tendency to respond to one-half of the stimuli could be explained in the following way. Since all the patterns are biased towards the template, the more the proportion of stimuli for which a novelty neuron responds is greater than one-half, the more anti-correlated its weights become to the template (due to anti-Hebbian learning). Hence the next stimulus is more likely to be anti-correlated to the weights of the neuron, and the membrane potential for this stimulus is likely to be lower; hence the neuron is less likely to be active. A similar argument can be applied if the coding is more sparse.

From Equations E.8 and E.14, it follows that the evolution of s_i^q may be described by the following one-dimensional nonhomogeneous random walk:

$$s_i^{q+1} = \begin{cases} s_i^q + 1 & with\ prob. \approx \frac{1}{2} - s_i^q B \\ s_i^q - 1 & with\ prob. \approx \frac{1}{2} + s_i^q B \end{cases}, \quad \text{where}$$

$$B = b^2 \sqrt{\frac{N}{2\pi q}} \qquad \text{(E.15)}$$

In Equation E.15, B describes the strength of the bias towards 0; i.e., the larger B, the more the value of s_i^q is being attracted towards 0. Numerical simulation shows that after large numbers of iterations, the average position in the walk is 0 (i.e., $\langle s_i^q \rangle = 0$), and the square of the average distance to 0 is equal to 1/4B, i.e.,

$$\langle (s_i^q)^2 \rangle = \frac{1}{4B} \qquad \text{(E.16)}$$

The above properties of the random walk of Equation E.15 may also be shown analytically, by approximating this discrete random walk by the following stochastic differential equation:

$$\dot{s}_i = -s_i B + \varepsilon \eta(t) \qquad \text{(E.17)}$$

In Equation E.17, $\eta(t)$ denotes [white] noise and ϵ denotes the amplitude of the noise. Holmes et al. (1996) showed that for Equation E.17 when $t \to \infty$, the probability distribution of the values of s_i becomes a normal distribution with mean 0 and variance:

$$\langle (s_i)^2 \rangle = \frac{\varepsilon^2}{2B} \qquad \text{(E.18)}$$

Both Equations E.16 and E.18 show that the variance of s_i is inversely proportional to B. Equations E.16 and E.18 differ only by a numerical constant, i.e., by a factor of $2\epsilon^2$. Because the calculations of this Appendix show that the proportionality is of greater significance than the absolute value, this difference will be disregarded.

Using Equations E.16 and E.15, we may approximate:

$$s_i^q \approx \theta\left(0,\ \sqrt{\frac{1}{4B}}\right) = \theta\left(0,\ \sqrt{\frac{1}{4b^2}\sqrt{\frac{2\pi q}{N}}}\right) \qquad \text{(E.19)}$$

Ignoring numerical constants (as assumed in the beginning of the Appendix) and from the definition of Equation E.7, we get

$$D^2\left(\sum_{\mu=2}^{P} y_i^\mu\right) \approx D^2(s_i^P) \approx \frac{1}{b^2}\sqrt{\frac{P}{N}} \qquad \text{(E.20)}$$

From Equations E.6 and E.20, we can approximate the distribution of the two inner summations in the noise term (Equation E.4):

$$\sum_{\mu=2}^{P} y_i^\mu \frac{1}{N} \sum_{j=1}^{N} x_j^1 x_j^\mu \approx \theta\left(0,\ \sqrt{\frac{P}{N} + \frac{1}{b^2}\sqrt{\frac{P}{N}\, b^4}}\right) \qquad \text{(E.21)}$$

Hence the noise term may be approximated by:

$$Noise \approx \theta(0,\ \sqrt{P + b^2\sqrt{PN}}) \qquad \text{(E.22)}$$

To summarise, we have calculated that the average value of the decision function for familiar patterns is lower by about N than for novel patterns, and the variance of the decision function is equal to

$$P + b^2\sqrt{PN}$$

Hence d (Equation A.7) may be used as the familiarity criterion, and the middle value (between the mean d for novel and familiar patterns) may be used as a familiarity discrimination threshold.

As for the Hebbian model (Appendix C), we consider the network as working well if the probability of error is less than 1%. An error occurs if the noise is higher than the threshold. To calculate the maximum acceptable number of stored patterns P_{max}, we must solve the following equation:

$$\Pr\left[\theta(0,\ \sqrt{P_{max} + b^2\sqrt{P_{max}N}}) < \frac{N}{2}\right] = 0.99 \qquad \text{(E.23)}$$

Checking the value of the cumulative standard normal distribution for 0.99 as in Appendix C, and ignoring numerical constants we get

$$P_{max} + b^2\sqrt{P_{max}N} \approx N^2 \qquad \text{(E.24)}$$

To solve Equation E.24 we transform it to a polynomial form:

$$P_{max}^2 + P_{max}(-2N^2 - Nb^4) + N^4 \approx 0 \qquad \text{(E.25)}$$

We use a standard technique to solve quadratic Equation E.25:

$$\Delta = 4N^3b^4 + N^2b^8 \qquad \text{(E.26)}$$

The second term in Equation E.26 is much smaller than the first; hence we discard it. Now we solve Equation E.25 and obtain

$$P_{max} \approx \frac{2N^2 + Nb^4 - \sqrt{4N^3b^4}}{2} \qquad \text{(E.27)}$$

Term Nb^4 in Equation E.27 is much smaller than the other terms. We discard it and obtain

$$P_{max} \approx N^2 - N^{3/2}b^2 \qquad \text{(E.28)}$$

As a result of approximations made during the derivation of capacity (in particular due to ignoring all the numerical constants), Equation E.28 shows only a qualitative relationship between capacity and the size of the network. To find the precise equation for

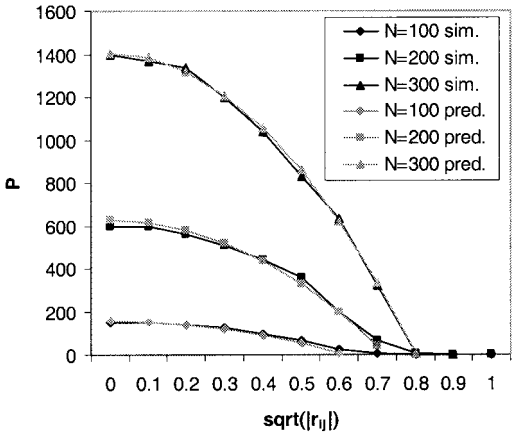

FIGURE 9. Capacity of the version of the anti-Hebbian model analysed in Appendix E. Methods of simulation and notation as in Fig. 5.

capacity, we fitted two numerical constants, one multiplying each of the terms in Equation E.28 to the results of simulations. First, we simulated a version of the anti-Hebbian model analysed here, where the weights are modified according to Equation E.1. The capacity estimated in simulations is shown in Figure 9. It can be described by the following equation (shown by the grey curves in Fig. 9):

$$P_{\max} \approx 0.0156N^2 - 0.42N^{3/2}b^2 \qquad \text{(E.29)}$$

Then we simulated the original anti-Hebbian model described in Appendix A.2. The capacity obtained in simulations is shown by the black curves in Figure 5d. The capacity of the original anti-Hebbian model (Fig. 5d) is slightly lower than the capacity of its version analysed here (Fig. 9) and may be approximated by the following equation (shown by the grey curves in Fig. 5d):

$$P_{\max} \approx 0.013N^2 - 0.31N^{3/2}b^2 \qquad \text{(E.30)}$$

Comparing Equations E.29 and E.30 demonstrates that using the learning rule of Equation A.6 instead of Equation E.1 changes only the values of the numerical constants but not the nature of relation between capacity and size of the network.

To summarise, Equation E.30 describes the capacity of the fully connected anti-Hebbian network for patterns biased towards a template. Note that when N grows, the second term in Equation E.30 becomes relatively small in comparison with the first term. Hence, for large networks, the capacity of the anti-Hebbian model converges to being proportional to the number of synapses in the network, as for uncorrelated input patterns. Thus, this Appendix shows that the anti-Hebbian model is very robust to the correlation in the responses of the input neurons. Correspondingly, for correlated input patterns, the anti-Hebbian model achieves much larger capacity than the Hebbian model.

APPENDIX F. ABILITY OF THE HEBBIAN MODEL TO DETECT UNUSUAL STIMULI

This Appendix calculates the ability of the Hebbian model to discriminate whether a pattern comes from the set of patterns with correlation matrix **R** or is an unusual type of pattern from another set with matrix **R′**. We showed in Appendix C that the average value of d after presentation of a novel pattern coming from a distribution described by **R** is equal to NPr^2. After presentation of a novel pattern x′ coming from a distribution described by **R′**, d is equal to

$$d(x') = \frac{1}{N} \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \sum_{\mu=2}^{P} x'_i x'_j x_i^{\mu} x_j^{\mu} \qquad \text{(F.1)}$$

The average of function d in Equation F.1 is equal to $NP\langle r_{ij} r'_{ij} \rangle_{i \neq j}$, and the variance depends on **R** and **R′**. For simplicity, let us consider the case when **R′** is the identity matrix; i.e., patterns x′ are not correlated. The average of d(x′) is then 0; hence a pattern may be classified as coming from a distribution described by **R′** if $d < NPr^2/2$. Analysing the square of d(x′) and then using similar techniques to those of the previous Appendices, the probability of correct discrimination is equal to

$$\Pr_{OK} = \frac{1}{2}\Phi\left(\frac{r^2}{4}\sqrt{\frac{N}{r^3 - (r^2)^2}}\right) + \frac{1}{2}\Phi\left(\frac{N}{4}\sqrt{r^2}\right) \qquad \text{(F.2)}$$

The first term of Equation F.2 corresponds to the probability of correct classification for patterns coming from the distribution described by **R**, and the second term to that from patterns coming from the distribution described by **R′**. Note that when r^2 is different from 0, and N increases, the arguments of both normal cumulative distributions Φ increase; hence the values of Φ converge to 1, and the probability of correct discrimination converges to 1 as well.

To summarise, this Appendix shows that the Hebbian model can detect unusual input patterns with a probability of error that rapidly converges to 0 for larger networks. Other familiarity discrimination networks also have this ability as demonstrated in Figure 8.