TOWARDS A COMPETITIVE LEARNING MODEL OF MIRROR EFFECTS IN YES/NO RECOGNITION MEMORY TESTS

K. C. DIETZ*, H. BOWMAN and J. C. VAN HOOFF

Centre for Cognitive Neuroscience and Cognitive Systems, University of Kent,
Canterbury, CT2 7NF,UK
**E-mail: kcd5@kent.ac.uk*
www.kent.ac.uk

Manipulations of encoding strength and stimulus class can lead to a simultaneous increase in hits and decrease in false alarms for a given condition in a yes/no recognition memory test. Based on signal detection theory, the strength-based 'mirror effect' is thought to involve a shift in response criterion/threshold (Type I), whereas the stimulus class effect derives from a specific ordering of the memory strength signals for presented items (Type II). We implemented both suggested mechanisms in a simple, competitive feed-forward neural network model with a learning rule related to Bayesian inference. In a single-process approach to recognition, the underlying decision axis as well as the response criteria/thresholds were derived from network activation. Initial results replicated findings in the literature and are a first step towards a more neurally explicit model of mirror effects in recognition memory tests.

1. Introduction

The accommodation of mirror effects in recognition tests has long posed a puzzle for memory researchers and has caused them to revise their assumptions of the underlying decision process (for a review, see Ref. 1). In a typical verbal yes/no recognition test, a list of individual words is presented during a study phase. In a subsequent test phase, the studied old words are randomly interleaved with new, not previously presented, words. For each test item, participants have to give a response, indicating whether they think it is old ('yes' response) or new ('no' response). The resulting decision matrix contains four possible outcomes: hits ('yes' responses to old items), misses ('no' responses to old items), false alarms ('yes' responses to new items) and correct rejections ('no' responses to new items).

A mirror effect occurs when there are two conditions that differ in their ease of recognition, and the easier condition shows not only a higher hit rate than the harder condition, but (perhaps surprisingly) also a lower false alarm rate.[2] While the generality of this effect has been questioned (*e.g.* Ref. 3), it is generally accepted as a 'regularity of recognition memory' (*e.g.* Ref. 4, p.177).

We introduce theories that explain which mechanisms may give rise to mirror effects, and describe a preliminary neural network model implementing these in a biologically plausible way.

2. Signal detection theory and the mirror effect

Conventionally, recognition decisions are analyzed from a signal detection perspective (for a detailed introduction, see Ref. 5). Two underlying factors are assumed: the strength of the memory signal elicited by a test item[a], and its relation to the placement of the participant's response criterion/threshold.

Memory signals for old and new items are represented as two Gaussian distributions of unequal variance.[6] The variance of the old distribution exceeds that of the new distribution, as it reflects the variability of learning in the study phase in addition to noise.[7] The distance between the means of the old and new distributions (in units of standard deviation) determines the ease of discrimination and is termed d'.

The criterion/threshold can be thought of as a single point along the strength-of-evidence axis. Test items whose memory signal exceeds this criterion/threshold value receive a 'yes' response, resulting in hits for old items and false alarms for new items. An optimal decision criterion/threshold maximizes correct responses and would be placed at the point of intersection of the old and new distributions.

Based on this signal detection framework, it is thought that two mechanisms can give rise to a mirror effect: shifts in the absolute placement of the response criterion/threshold along the strength-of-evidence axis (Type I) and changes in the underlying distributions (Type II).[8]

2.1. *Type I mirror effects*

Type I mirror effects (see Fig. 1) are usually observed for strength manipulations of otherwise identical stimulus materials.[9] For example, repeating

[a] Whether the signal is based on a single continuous variable, combines a continuous with a dichotomous variable or involves a second independent process, is beyond the scope of this paper.

half of the items in the study phase will lead to more hits and fewer false alarms compared to items presented only once.

Repetition has long been shown to lead to more accurate and robust encoding (*e.g.* Ref. 10), so that the mean of the strong old condition (repeat presentation) tends to be located further along the decision axis (further to the right) than that of the weak old condition (single presentation). This is also reflected by a larger d' value in the strong condition.

Assuming the mean signal of the new distribution remains constant, participants have to shift their decision criterion/threshold upwards (to the right) to maintain an optimal response strategy. This criterion/threshold shift accounts for fewer false alarms in the strong condition, while the simultaneous upwards (to the right) movement of the old distribution explains the increased hit rate. Note that there is only a single new distribution in a Type I mirror effect.

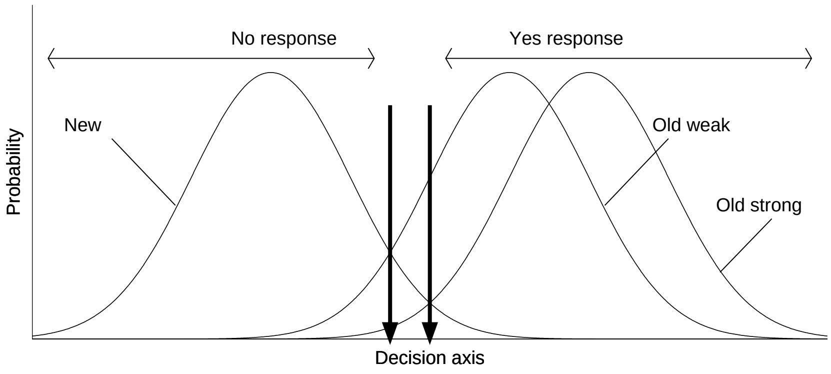

Fig. 1. Type I mirror effect: Underlying distributions and response criteria/thresholds.

2.2. *Type II mirror effects*

Type II mirror effects (see Fig. 1) are usually observed for manipulations of stimulus class, for example where high- and low frequency words are presented during both study and test. In this case, no criterion/threshold shift is observed even if explicit cues are provided about which items are of high- and low frequency;[8] yet the low frequency words consistently produce lower false alarm- and higher hit-rates. Given that low frequency words tend to have fewer definitions and are used in less varied contexts, they are thought to elicit a lower strength memory signal than high frequency items when new. By the same token, they are also more accurately encoded, explaining the advantage for recognition of low frequency old items.[11,12] It has also been proposed that due to their comparative 'novelty', low frequency words elicit increased attention and therefore more elaborative processing.[2,13] As a result of the increased separation of the old and new distribution for low-frequency words, d' is larger than for high-frequency words.

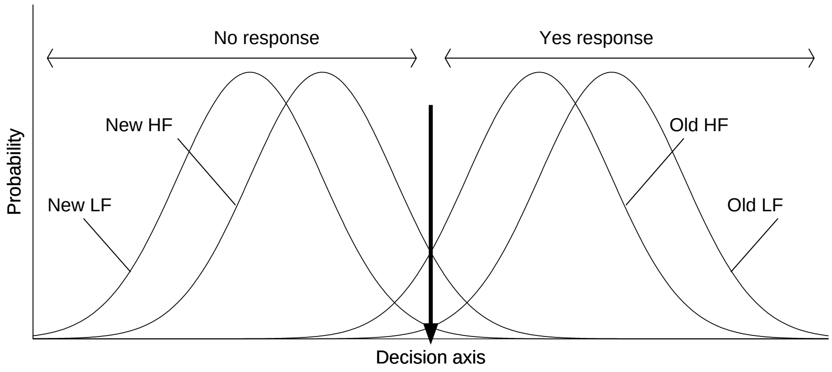

Fig. 2. Type II mirror effect: Underlying distributions and response criterion/threshold. LF = low frequency words, HF = high frequency words.

3. Our model

Although there are a large variety of single- and dual-process models of recognition (*e.g.* Ref. 14,15, for a review, see Ref. 7), not all address mirror effects. Those that do are often single-process Bayesian models[9,12,16,17] (but see *e.g.* Ref. 18 for an exception). While theoretically pleasing, such Bayesian models largely ignore issues of neural implementation by using mathematical quantities without specifying how these might be calculated by the brain. In this paper, we present a first step towards a neurally more detailed model of the mirror effect in yes/no recognition memory tests.

We use McClelland and Chappell's subjective likelihood model[16] as a starting point, as it reproduces a wide variety of recognition memory phenomena. As in their model, each simulation run emulates a yes/no recognition test in which a participant is presented, one 'word' at a time, with a single, multi-item study list. Study items are learned, that is encoded into memory. In a subsequent test phase, (studied) old items are mixed with new, not previously presented, items. No learning occurs at test, but stored information about each presented item is retrieved and a yes/no decision is made according to its position relative to the response criterion/threshold.

We preserved a number of the implementation ideas but deviate from their key idea that the recognition decisions are based on a likelihood evaluation: in our model, the decision axis represents simple memory trace strength often termed 'familiarity'.

The model is a simple two-layer feed-forward network with a competitive Conditional Principal Component Analysis learning rule (CPCA, Ref. 19) using Winner-Takes-All (WTA). We use distributed representation of items on the input layer, but a localist representation on the output layer. Each item is a binary vector of 'features' (such as orthographic properties or semantic and contextual associations) across the 500 units of the input layer. Fifty randomly chosen features are active (1), all others are inactive (0), with the exception of low frequency items, which have one fewer active features. This reflects the previously mentioned property that low frequency items have fewer definitions and appear in fewer contexts. Each stimulus class comprises 30 such patterns. Initially, the output layer consists of 120 detectors. These are reduced to a maximum number of 60 (for 2 classes of 30 items) after learning, so that invariably, one detector comes to encode one item presented during the study phase (old items). The exact number is subject to constraints of the learning algorithm, which allows the possibility that a single detector comes to encode multiple items, but the initial weight settings keep the probability of this low.

In the study phase, items are presented in random order. In the simulation of a Type I mirror effect (strength-based), half of the items at study are presented twice whilst all other parameters are kept constant. In the simulation of a Type II mirror effect (frequency-based), half the presented patterns are of low frequency, with one less active input unit and a higher learning rate. The latter reflects the previously discussed increased attention to low frequency items. Weights between the input and output layer are initialized to random values in the range [0.45–0.55], which is the initial conditional probability that a given input unit is active for a specific detected item. When an input pattern is presented, competitive winner-takes-all learning takes place. The activation of each detector (denoted y_j) is calculated by feeding its summed, weighted and normalized net input (denoted η_j, Eq. 1) through a sigmoid function with a gain term $\lambda = 10$ and bias term $\beta = 0.5$ (Eq. 2), where N is the number of active units in an input pattern.

$$\eta_j = \frac{1}{N} \sum_i x_i \; w_{ij} \tag{1}$$

At this stage, Gaussian noise is added $G \sim N(0, 0.03^2)$.

$$y_j = G + \frac{1}{1 + e^{\lambda(-\eta_j + \beta)}} \tag{2}$$

Weights of the most active detector are adjusted using a CPCA Hebbian learning rule (Eq. 3, Ref. 19), bounded between [0–1]. Weights between active input units and detectors are increased and connections between active detectors and inactive input units are decreased. The rate of change is determined by the learning constant ϵ. (No change occurs for weights to inactive detectors, unlike for some other biologically plausible Hebbian learning rules.[b])

$$\Delta w_{ij} = \epsilon \; y_j (x_i - w_{ij}) \tag{3}$$

For each trial (that is, for each presented item in the study phase), the noisy activation of the winning detector is added to a weighted average (denoted avg_{max}). Relative proportions are determined by the time constant τ, which was set to 0.7, so that the current trial contributes 0.3 and the previous average 0.7.

$$avg_{max}(t) = \tau \; avg_{max} \; (t-1) + (1-\tau) \; y_j \tag{4}$$

In Equation 4, y_j denotes the activation of the winning detector and t indexes trials, *i.e.* patterns being presented to the network. This 'time averaging' is a simple and efficient method used by biological neurons for increasing the signal-to-noise ratio.[20] Initially, all time averages have a value of 0.5 to indicate that no information is known about the relationship between the detector activation and the input feature activation. In the Type I mirror effect simulations, time averages for weak *old* and strong *old* items are calculated throughout the study phase. These are based on the noisy activation value of the most active detector for a given pattern. In the Type II mirror effect simulation, a single time average for *old* items is calculated by collapsing across high and low frequency words.

On completion of the study phase, but before the test phase, thirty random *new* patterns are generated per condition (Type I simulation: new, Type II simulation: weak class new, strong class new). These are presented

[b] We would like to thank Max Garagnani for pointing this out.

Table 1. Model generated data for Type I and Type II mirror effects over 100 simulations. Hits and false alarms (FA) are shown in percentages. d' values reflect the ease of discrimination. Criteria/thresholds are based on *estimated* activation averages.

	Weak cond.	Strong cond.	High freq.	Low freq.
Hit	75.100	91.767	80.767	96.400
FA	7.000	2.900	16.400	1.1
d'	2.153	3.285	1.847	4.089
Criterion/Threshold	.541	.558	0.550	

to the network using the fixed learned weights from the study phase. Time averages are calculated based on network activation as before (see Eq. 4). For the Type II simulation, the time average is collapsed across high and low frequency words to generate a single time average for *new* items.

In the test phase, previously presented items are mixed with an equal number of new items and presented in random order. Activation values are calculated as before. For simplicity, we assume an unbiased response criterion/threshold in each simulation, located half-way between the *estimated* average activation of the maximally active detector for old and new items (*i.e.* avg_{max}), both derived prior to the test phase (see Eq. 4). The criterion/threshold is based on *estimates* rather than actual values, as participants are assumed not to have access to this information.[21]

In the Type I mirror effect simulation (strength-based), two criteria/thresholds are used: one between new and weak old items (single presentation) and one between new and strong old items (repeat presentation). This represents the notion that on average, participants have a higher feeling of familiarity for repeated items and thus require a higher activation level before declaring these to be old. In contrast, weak old items elicited a lower level of memory activation, so that a less stringent criterion/threshold is adopted.[8]

In the Type II mirror effect simulation (stimulus class based), a single decision criterion/threshold is calculated, which collapses new and old items across low- and high frequency. This is based on the observation that participants do not appreciate that low frequency items are more memorable, but instead adopt a single response criterion/threshold, even if provided with explicit cues about the item type.[8]

4. Results

Results for 100 simulation runs of Type I and Type II mirror effects are given in Table 1 and depicted in Fig. 3 and Fig. 4. Distributions are plotted for the noisy activation values for the most active detector by item class. As previously suggested, criteria/thresholds were placed in an unbiased 'optimal' way, half-way between the *estimated* activation means for old and new items.

4.1. *Type I mirror effect*

In the Type I mirror effect simulation (strength-based), strong and weak old items differed only by the number of repetitions during the study/learning phase. Weak items were presented once, strong items twice, both with a learning constant $\epsilon = .05$. In line with empirical data, new items, whose representation did not differ, form a single distribution (see Fig. 3).

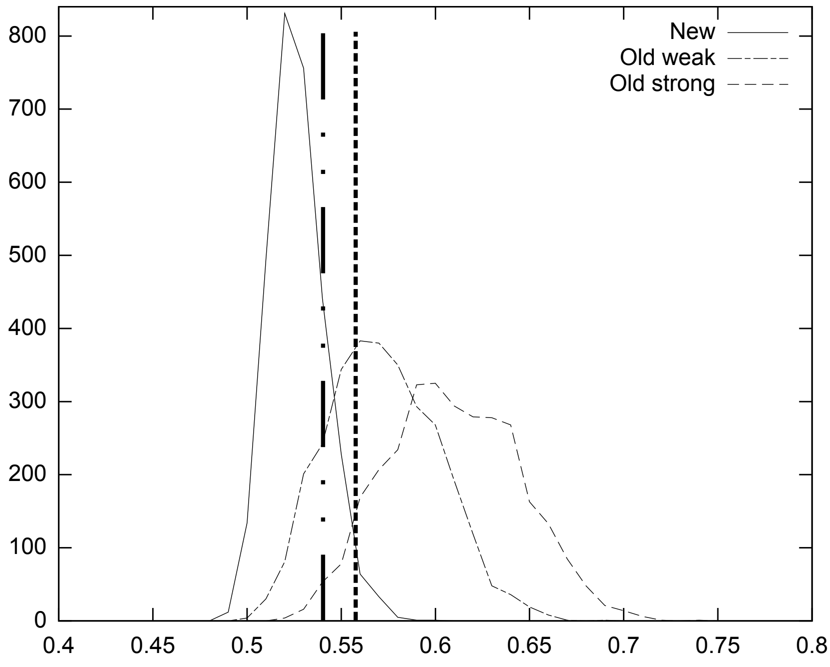

Fig. 3. Type I mirror effect simulations: Frequency (y-axis) distributions for actual maximal activation values (x-axis) per item type, based on 100 experimental runs. Response criteria/thresholds for weak old (dash-dot line) and strong old items (short-dashed line) are placed mid-way between new and old *estimated* maximal activation averages.

Distributions of weak and strong old items separate due to more accurate encoding of the latter, resulting in a larger d' in the strong condition.

Old items have a larger variance than new items as they were also subject to variability during learning. The response criterion/threshold in the strong condition has shifted upward as a function of the higher estimated mean activation of strong old items. Hits and false alarms follow a mirror effect pattern (see Table 1).

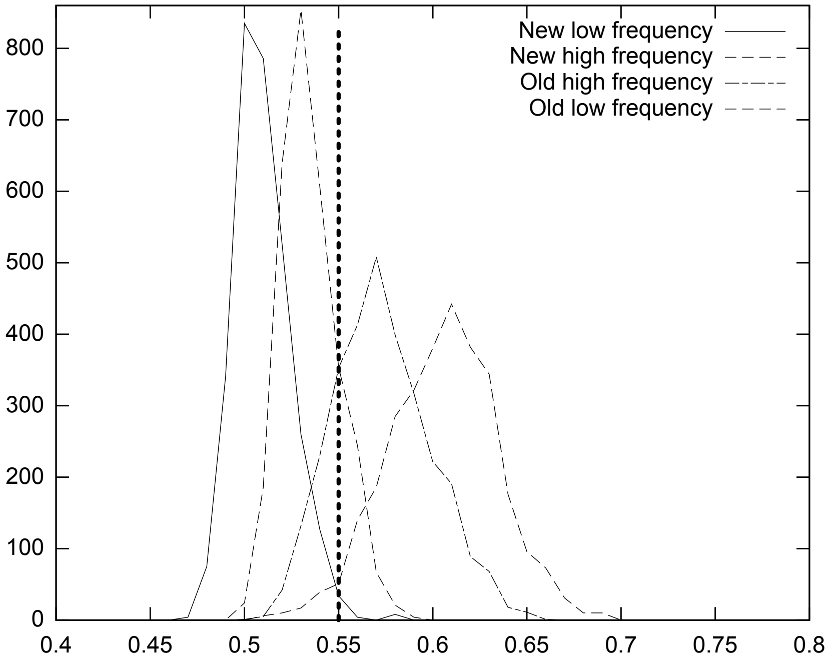

Fig. 4. Type II mirror effect simulations: Frequency (y-axis) distributions for actual maximal activation values (x-axis) per item type, based on 100 experimental runs. A single response criterion/threshold (short-dashed line) is used for high- and low frequency items and placed mid-way between collapsed new and old *estimated* maximal activation averages.

4.2. *Type II mirror effect*

In the Type II mirror effect (stimulus class based) simulation, low-frequency differed from high frequency items in two respects. They had one less active unit (49 compared to 50), which was meant to reflect the fact that low-frequency words tend to have fewer definitions and are thus used in less varied contexts.[11,12] They also had a higher learning rate η than high frequency items (0.09 compared to 0.04), simulating better encoding due to increased attention to their comparative 'novelty'.[2,13] Based on Stretch and Wixted's findings,[8] a single criterion/threshold was used for high- and low frequency items. The model reproduces a Type II mirror effect, with appropriate hits, false alarms and d' values (see Fig. 4 and Table 1).

5. Conclusions and further work

We have presented a neural network model of memory processes underlying yes/no decisions in a recognition memory test, which reproduces Type I and Type II mirror effects. The approach takes inspiration from Bayesian models of recognition memory, especially from McClelland and Chappell's subjective likelihood model.[16]

One of the most significant differences between the approach presented here and existing Bayesian models[8,12,16,17] is that we obtain a mirror effect without an explicit likelihood ratio calculation. We have shown here that the two basic classes of mirror effects can be generated from a simple, competitive learning neural network in which the values for the familiarity axis are directly generated from the activation of the winning detectors. The simplicity of this approach, which is based on a signal detection, single-process view of recognition memory, along with the direct calculation of the criterion/threshold from neural activation are the key benefits of the model we have introduced.

A limitation of our model concerns the treatment of low frequency words. In order to replicate Type II mirror effects, we had to combine two assumptions. Firstly, low frequency words elicited increased amounts of attention compared to high frequency words,[2,13] which was reflected by their higher learning rate. Secondly, low frequency words tend to have fewer definitions and are associated with fewer contexts, which was reflected by one less active input unit or 'feature' for this stimulus class.[11,12] Some competitor models distinguish low- and high frequency through just one manipulation.[2,11–13]

We further assume that neurons can perform a maximum-like operation (which results in an output signal that approximates the maximum among several input signals) to calculate response citeria/thresholds. This operation has been shown to be approximated by complex cells in the visual cortex of cats and neurons in Area V4 of macaques[22] and demonstrated in a neurophysiologically plausible way for feed-forward models.[23]

We also assume that humans are able to distinguish high-strength from low-strength stimuli, yet we do not implement how this might be achieved.

Whilst this is not ideal, this assumption is common in the literature (*e.g.* Ref. 8).

Our model shares some similarities with Bogacz, Brown and Giraud-Carrier's familiarity discrimination model.[24] This model closely reproduces observed activation patterns of 'novelty' neurons in the perirhinal cortex with a two-phase three-layer network (binary input, familiarity discrimination, decision) with primarily feed-forward connections, biologically plausible parameters and Hebbian learning rules. While the authors did not use sparse coding, they demonstrated by simulation that the network's behaviour would essentially remain unchanged.

During the critical initial period (the familiarity discrimination phase), the Bogacz et al.[24] model's familiarity discrimination neurons (FDNs) are likely to become more active for familiar patterns than for novel ones. This is because Hebbian learning leads the synaptic weights to reflect the correlation between the active inputs for a given FDN. The number of FDNs that can become active for a given input are limited by fixed high synaptic weights between specific input and output units in combination with inhibition. Our implementation of a CPCA learning rule in combination with winner-takes-all achieves similar results, although we do not use homosynaptic long-term depression.

In comparison to Bogacz et al.,[24] we are less explicit in relating our model parameters back to functional neuroanatomy and we do not analyse our model storage capacity mathematically. However, the focus on modelling mirror effects and the use of a signal detection framework distinguishes our work.

For further work, we would like to use a principled approach (such as maximum likelihood estimation) for the setting of free parameters and extend the model to generate reaction time data. We want to move from qualitative to quantitative modeling. We aim to generalize the model to cases in which mirror effects are predicted but not observed and eventually to a broader range of recognition memory phenomena (*e.g.* the list length effect). Eventually, the model could be refined and extended to emulate known properties of functional neuroanatomy, like, for example, Norman and O'Reilly's model.[15]

Acknowledgements

We would like to thank our anonymous reviewers for their comments which helped to improve this paper.

References

1. R. Ratcliff and G. McKoon, *Oxford Handbook of Memory* (OUP: New York, 2000), ch. Memory models, pp. 571–582.
2. M. Glanzer and J. K. Adams, *Memory and Cognition* , 8 (1985).
3. R. L. Greene, *The Foundations of Remembering: Essays in Honor of Henry L. Roediger, III* (Psychology Press: Hove, UK., 2007), ch. Foxes, Hedgehogs, and Mirror Effects: The Role of General Principles in Memory Research, pp. 53–66.
4. G. Stenberg, M. Johansson and I. Rosén, *Acta Psychologica* , 174 (2006).
5. N. A. Macmillan and C. D. Creelman, *Detection Theory: A user's guide*, 2nd edn. (Psychology Press: Hove, UK., 2005).
6. R. Ratcliff, C. F. Sheu and S. D. Gronlund, *Psychological Review* **99**, 518 (1992).
7. J. T. Wixted, *Psychological Review* **114**, 152 (2007).
8. V. Stretch and J. T. Wixted, *Journal of Experimental Psychology: Learning, Memory and Cognition* **24**, 1379 (1998).
9. M. Glanzer, J. K. Adams, G. J. Iverson and K. Kim, *Psychological Review* **100**, 546 (1993).
10. A. M. Glenberg, *Memory and Cognition* **7**, 95 (1979).
11. M. Glanzer and N. Bowles, *Journal of Experimental Psychology: Human Learning and Memory* **2**, 21 (1976).
12. R. M. Shiffrin and M. Steyvers, *Psychonomic Bulletin and Review* **4**, 145 (1997).
13. M. Glanzer and J. K. Adams, *Journal of Experimental Psychology: Learning, Memory and Cognition* **16**, 5 (1990).
14. R. A. Diana, A. P. Yonelinas and C. Ranganath, *Trends in Cognitive Sciences* **11**, 379 (2007).
15. K. A. Norman and R. C. O'Reilly, *Psychological Review* **110**, 611 (2003).
16. J. L. McClelland and M. Chappell, *Psychological Review* **105**, 724 (1998).
17. M. Murdock, *Psychonomic Bulletin and Review* **10**, 570 (2003).
18. M. Cary and L. M. Reder, *Journal of Memory and Language* **49**, 231 (2003).
19. R. C. O'Reilly and Y. Munakata, *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain* (Cambridge: MIT Press, 2000).
20. J. L. McClelland, *Psychological Review* **86**, 287 (1979).
21. D. L. Hintzman, *Journal of Experimental Psychology: Learning, Memory and Cognition* **20**, 201 (1994).
22. M. A. Giese and T. Poggio, *Nature Reviews Neuroscience* **4**, 179 (2003).
23. T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman and T. Poggio, *A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex*, tech. rep., MIT: Cambridge, MA (2005), CBCL Paper 259/AI Memo 2005-036.
24. R. Bogacz, M. W. Brown and C. Giraud-Carrier, *Journal of Computational Neuroscience* **10**, 5 (2001).