Modeling Hippocampal and Neocortical Contributions to Recognition Memory

Kenneth A. Norman, Randall C. O'Reilly, & David E. Huber
Dept. of Psychology, University of Colorado at Boulder

Computational Cognitive Neuroscience Laboratory

poster presented at the 2000 Cognitive Neuroscience Society meeting

questions/comments? email Ken Norman, norman@psych.colorado.edu

Introduction

In this poster, we present a biologically-based dual-process model of recognition memory.

Dual-process models posit that recognition judgments are based on:
- recollection of specific details, and
- nonspecific feelings of familiarity

Recollection depends on the hippocampus.

Recent data suggest that medial temporal neocortical regions (*MTLC*) play an important role in supporting familiarity-based recognition (for a review, see Aggleton & Brown, 1999).

We seek to understand, in mechanistic detail, how MTLC and the hippocampus contribute to recognition memory, by constructing neural network models of these structures, and using them to simulate recognition data from lesioned and intact subjects.

Two Incompatible Goals

Our overall view of neocortical and hippocampal processing builds on the *Complementary Learning Systems Framework* set forth by McClelland, McNaughton, & O'Reilly (1995).

This framework starts from the premise that learning about *specifics* and extracting *generalities* are computationally incompatbile tasks. Thus, we have evolved specialized networks for performing these tasks.

Neocortex learns slowly, integrating across episodes to arrive at a representation of what is *generally* true in the environment.

Hippocampus learns rapidly, binding together co-active cortical neurons in a manner that allows stored patterns to be *completed* based on partial cues.

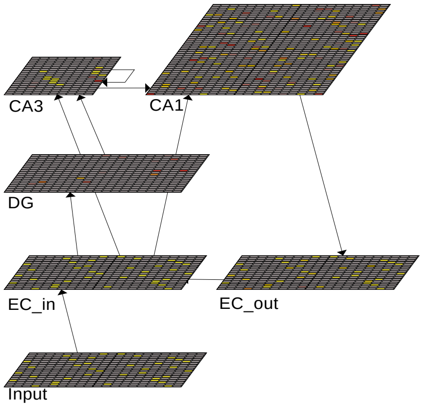
Neocortex assigns similar representations to similar stimuli => this allows it to *generalize* to novel stimuli based on their similarity to previously encountered stimuli

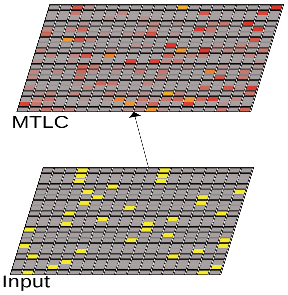
Hippocampus assigns distinct representations to stimuli => use of relatively non-overlapping ("pattern-separated") representations allows the hippocampus to learn rapidly without suffering from catastrophic interference

The Models

Both the hippocampal and neocortical networks were constructed using the Leabra model (O'Reilly & Munakata, 2000), which brings together several widely-accepted characteristics of learning in the brain -- including, but not limited to, Hebbian LTP/LTD (long term potentiation/ depression) and inhibitory competition between neurons.

The hippocampal network links overlapping patterns in entorhinal cortex (EC) to relatively non-overlapping patterns in region CA3; recurrent connections in CA3 bind together all of the units involved in representing a particular EC pattern; the CA3 representation is linked back to EC via region CA1.





The cortical network is a simple two-layer net, in which the hidden layer (corresponding to MTLC) encodes regularities that are present in the input layer (corresponding to "lower" cortical regions).

Initial simulations were run separately in the hippocampal and neocortical networks.

- The same inputs were presented to both models.
- The same parameter values were used for all of these simulations.

Recognition in the Hippocampus

Recognition in the hippocampal model is based on the extent to which the test cue is recalled. Specifically, we use the measure:

(# of recalled features that *match* the test cue) -
(# of recalled features that *mismatch* the test cue)

The mismatch term reflects the fact that lures sometimes trigger recollection of a similar studied item, and this can serve as grounds for rejecting the lure.

Recognition in MTLC: Familiarity as Sharpening

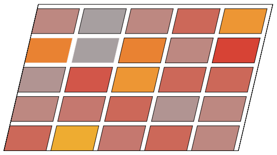
How can MTLC contribute to recognition after a single study exposure, if it is supposed to learn slowly (integrating over events)?

Although each weight changes only slightly when an item is studied, these small weight changes combine to yield a reliable, detectable effect on how that item is represented in MTLC.

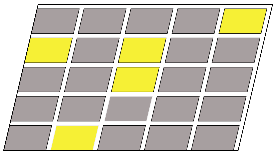
As items become more familiar, representations become *sharper*:

- *unfamiliar stimuli* ***weakly*** *activate a* ***large*** *number of units*

- *familiar stimuli* ***strongly*** *activate a relatively* ***small*** *number of units*



unfamiliar



familiar

Sharpening occurs because Hebbian learning specifically tunes some units to represent the stimulus.

When the stimulus is presented again, the well-tuned units are strongly activated, and these units suppress other, less well-tuned units via inhibitory competition.

Converging evidence from single-cell recording: Stimulus familiarization causes some neocortical neurons to fire less to that stimulus, whereas other neurons -- those selected to represent the stimulus -- fire more (e.g., Rolls et al., 1989).

To index sharpness -- and thus familiarity -- we compute the following measure:
average activity of units that are active (with act > .01)

- When a stimulus is unfamiliar, weakly active units drag down the average
- With familiar stimuli, units that are active tend to be *strongly* active => therefore, the average activity of these units will be high

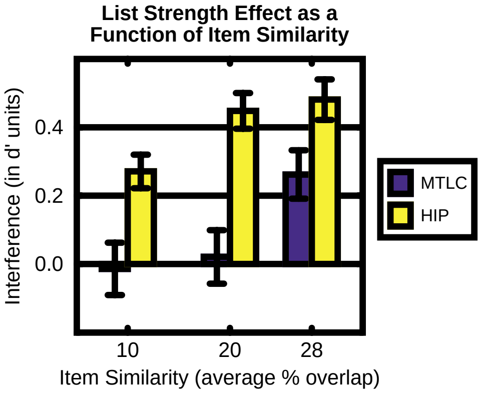
Interference: List Strength Effects

In both the hippocampal and cortical models, presenting an item multiple times improves recognition of that item (an *item strength* effect).

*Here, we explore whether there is also a **cost** associated with strengthening some memory traces: Does repeating some studied items impair recognition of other (non-repeated) studied items? This is called a **list strength** effect.*

We ran simulations examining the list strength effect; we also manipulated average *item similarity* to see if this factor interacts with list strength.

Data are presented using interference difference scores; positive values indicate the presence of a list strength effect. Error bars in this poster indicate 95% confidence intervals.



Key results:

- Hippocampus shows a list strength effect regardless of item similarity.

- MTLC only shows a list strength effect once item similarity exceeds 20% overlap.

Empirical support:

The model's prediction of a list strength effect for (hippocampal) recollection was confirmed by Norman (submitted).

The finding that overall recognition performance is typically *un*affected by list strength (e.g., Ratcliff, Clark, & Shiffrin, 1990) can be explained in terms of subjects relying primarily on neocortical familiarity.

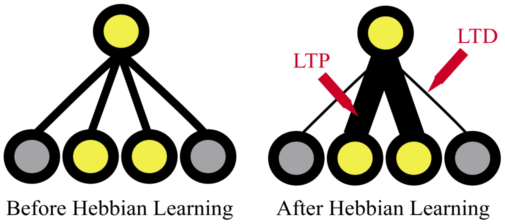
The model predicts that a recognition list strength effect should emerge when the contribution of hippocampal recollection is increased. Evidence consistent with this prediction was obtained by Norman (submitted).

Explaining the List Strength Results

*The lack of a list strength effect for MTLC (when item similarity is relatively low) can be explained in terms of **differentiation** -- strengthening an item's memory trace makes it less likely to be activated by other items (McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997).*

Differentiation occurs in the model because of Hebbian LTD.

Hebbian learning increases weights from active inputs to active MTLC units (LTP), but it also *decreases* weights from *in*active inputs to active MTLC units (LTD).



The LTD effect works to decrease interference, by decreasing the odds that an item's MTLC representation will be (spuriously) activated by other items at test.

When items are relatively dissimilar, LTD effects balance out LTP effects, and there is no interference.

*Interference occurs in the hippocampus because recollection is a competitive process (only one pattern can be recalled at a time), and because the hippocampus does not **completely** eliminate overlap between patterns.*

When a hippocampal unit is activated by two different inputs, strengthening the unit's connections to pattern X will necessarily lead to worse recall of pattern Y (since activated "X" features will compete with "Y" features).

Thus, even though there is *less* overlap between patterns in the hippocampus than in the cortex, this small amount of overlap is more *consequential* because of the "zero sum" nature of recollection.

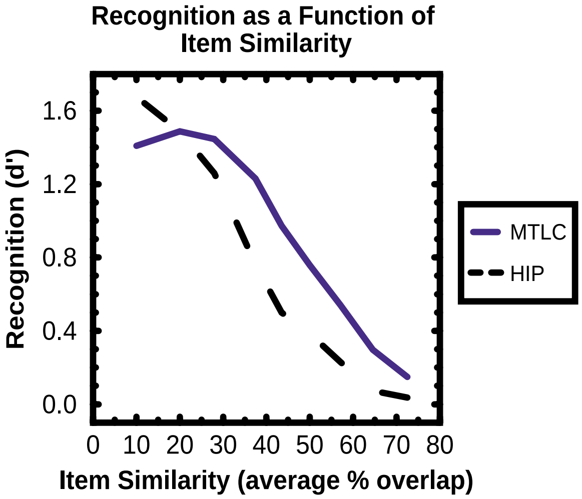
The fact that overlap is so harmful to recollection underlines the need for pattern separation in the hippocampus.

Interference:
Main Effects of Item Similarity

How does increasing the average similiarity of items to one another affect hippocampally-driven and neocortically-driven recognition?

Key result:

- *Moving from low to moderate levels of similarity, hippocampally-driven recognition* ***decreases*** *but MTLC shows a small but reliable* ***increase*** *in recognition performance.*



The decrease in hippocampally-driven recognition is not difficult to explain:

- Increasing similarity increases overlap in the hippocampus and (as discussed earlier) increasing overlap impairs recollection.

Why does increasing similarity improve recognition in MTLC?

- Increasing item similarity makes both studied items and lures more familiar.

- For complex reasons, the increase in familiarity is larger at first for studied items than lures, leading to an increase in d'.

Empirical support from studies of divided attention (at encoding):

=> Dividing attention at encoding should lower the distinctiveness of memory traces, by preventing subjects from elaborating on studied items.

- Studies using the process dissociation procedure (e.g., Jacoby, 1991) have found that dividing attention -- thus increasing trace similarity -- impairs recollection but not familiarity.

- Curran (in press) has isolated distinct ERP correlates of recollection and familiarity. In a recent study, Curran (personal communication) found that dividing attention at encoding led to an increase in the ERP effect associated with familiarity, and a decrease in the ERP effect associated with recollection.

ROC Curves

Does our model produce the ROC curves predicted by extant (non-mechanistic) dual-process models?

Yonelinas, Jacoby, and colleagues have devised several procedures for estimating recollection and familiarity from behavioral data (Jacoby et al., 1997); these procedures all rely on the following assumptions:

1) familiarity is an equal-variance signal detection process

2) recollection is a high-threshold process: Studied items are sometimes recollected (as having been studied), but this never happens for lures

3) recollection and familiarity are independent

Assumptions 1 and 2 lead to concrete, distinctive predictions about the shapes of ROC curves (generated by plotting hits vs. false alarms while varying response bias; Yonelinas et al., 1996):

- familiarity ROCs should be symmetrical and curvilinear

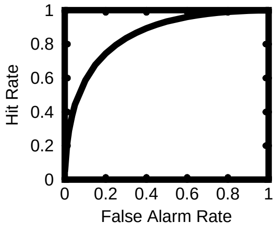
- recollection ROCs should be linear

Key result: The curves generated by our models are remarkably similar to the curves predicted by Yonelinas.

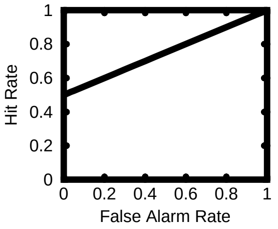
Our model therefore provides a principled justification for Yonelinas' assumptions 1 and 2 (assumption 3, the independence assumption, is addressed later in this poster...)

=> we should note, however, that our model does not behave *exactly* in accordance with Yonelinas' assumptions; for example, we treat recollection as a continuously varying signal (which can take on above-zero values for lures), whereas recollection is all-or-none in Yonelinas' framework

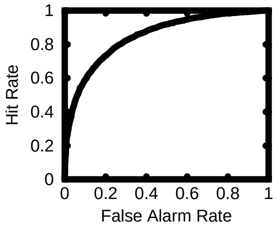
Idealized ROC for Familiarity



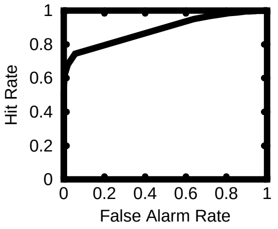
Idealized ROC for Recollection



ROC Curve for MTLC model



ROC Curve for HIP model



The high-threshold nature of recollection in our model (the fact that it is possible -- if you select a high enough threshold -- to eliminate recollection false alarms without totally eliminating recollection hits) is a consequence of hippocampal pattern separation:

- lures have to be quite similar to studied items before they trigger any kind of recollection

- when a lure is very similar to a studied item, it will often trigger recall of that item, and can be rejected based on mismatch between the recalled studied pattern and the test probe

By contrast, cortical processing is much more graded:

- the cortical familiarity signal triggered by a stimulus at test is a smoothly varying function of its similarity to studied items (i.e., its "global match")

- thus, lures that happen to share features with studied items will trigger a strong familiarity signal, resulting in false recognition

Lure Similarity and Test Format

*The ROC results suggest that MTLC should perform especially poorly, relative to the hippocampus, on recognition tests where lures are **related** to studied items.*

To review:

- related lures should trigger strong feelings of familiarity, leading to false recognition (and low d' scores) in the MTLC model; the hippocampus should show better discrimination because of its ability to pattern-separate similar inputs

- with *unrelated* lures, both models should show good discrimination (unrelated lures will be relatively unfamiliar, and they will not trigger recollection)

Effects of test format:

Although related lures will be highly familiar, they should be reliably *less* familiar than the corresponding studied item

=> *Thus, when lures are related to studied items, MTLC should greatly benefit from use of a **forced-choice** (FC) test procedure, in which subjects choose between studied items and corresponding lures*

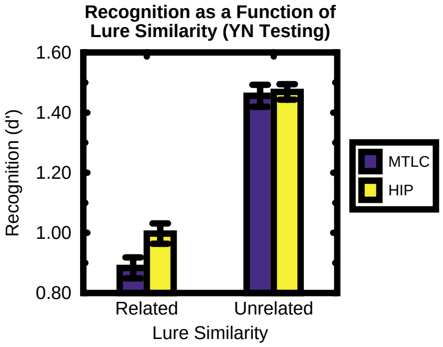
- this test format allows subjects to tune into the small but reliable familiarity differences that exist between these items.

We conducted simulations where we varied lure relatedness and test format: standard, yes/no (*YN*) single probe testing vs. FC testing. To facilitate comparison of the models, MTLC performance was matched to hippocampal performance in the unrelated lures condition.

Key result:

MTLC performed worse than the hippocampus on the YN related lures test, but MTLC was unimpaired on the FC related lures test.

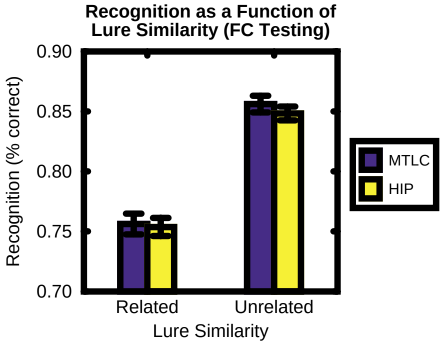
In summary, the models predict a 3-way interaction whereby MTLC performance is significantly worse (relative to the hippocampus) on YN tests with related lures, as compared to all other conditions defined by crossing test format (YN/FC) and lure relatedness.



Empirical support:

We tested the models' prediction via our collaboration with Andrew Mayes. Mayes' group has conducted extensive testing of patient YR, who suffered focal hippocampal damage, sparing surrounding MTLC regions.

YR is severely impaired at recalling specific details
=> thus, YR has to rely primarily on neocortical familiarity when making recognition judgments



Results are consistent with the predicted 3-way interaction!

- YR was significantly impaired on a YN recognition test which used related lures, but she performed slightly *better* than the control mean on a FC version of the same test (Holdstock et al., submitted).

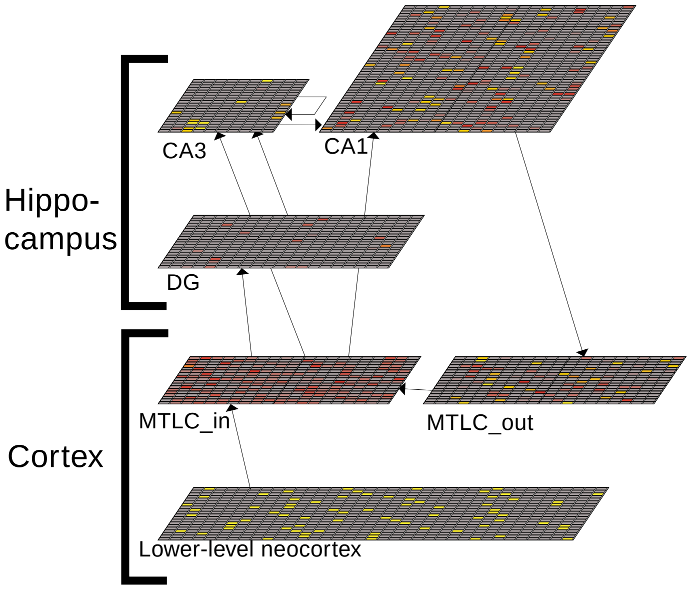
- YR was not significantly impaired on any of a large number of YN and FC recognition tests using unrelated lures.

- This pattern can not be explained in terms of difficulty confounds (e.g., controls found the YN related lure test to be easier than the FC related lure test).

The Combined Model and Independence

What is the statistical relationship between hippocampal recollection and MTLC familiarity? Specifically, are the two independent, as the process dissociation procedure assumes?

To address this question, we implemented a more realistic *combined* model, whereby the hidden (MTLC) layer of the cortical network serves as the input layer to the hippocampal network.-- this arrangement accurately reflects how the two structures are connected in the brain.



This model outputs a familiarity and recollection signal for each item. We ran a simple recognition simulation using the combined model and measured the extent to which the recollection and familiarity signals were correlated.

A priori, one might suspect that having the MTLC as input to the hippocampus would result in a correlation between these two systems.

Key result: Recollection and familiarity were ***completely independent****.*

This result shows that there are enough independent sources of variability in the two networks to eliminate correlations induced merely by the way the networks are connected.

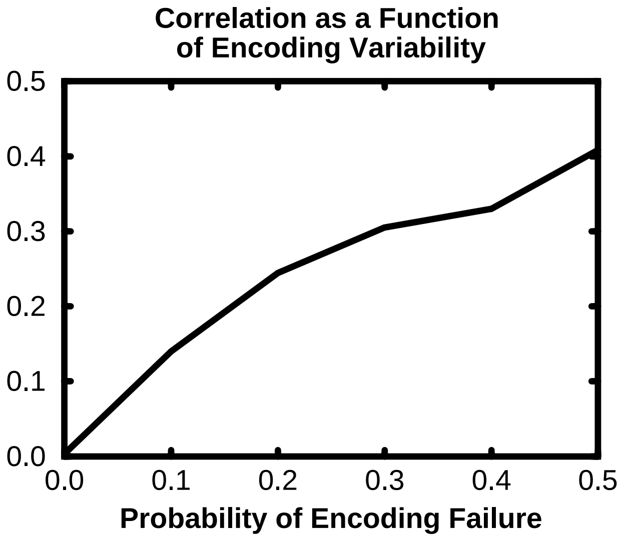
However, recollection and familiarity need not always be independent; the model can be used to explore when to expect independence violations.

For example, ***encoding variability*** *could produce dependence...*

=> if items vary substantially in how well they are encoded, poorly-encoded items will be unfamiliar and will not be recollected; well-encoded items will be more familiar, and more likely to trigger recollection.

We ran simulations in the combined model where we manipulated encoding variability by varying the probability of complete encoding failure from 0 to .50.

As predicted, increasing encoding variability increased the recollection-familiarity correlation.



Associative Recognition

A central feature of the hippocampus is that it forms conjunctive representations that bind together separate features of an event into a unitary representation.

*Is MTLC familiarity only sensitive to **individual features**, or is it also sensitive to **feature conjunctions**?*

We explored this using an associative recognition paradigm, in which subjects study pairs of stimuli (A-B, C-D); at test, subjects have to discriminate between studied pairs and *associative* lures generated by re-combining studied pairs (A-D, B-C).

The hippocampus can discriminate between studied items and associative lures based on recollection of actual study pairs (e.g., recollecting what was actually said in conjunction with "A" at study).

How will MTLC perform on associative recognition tests?

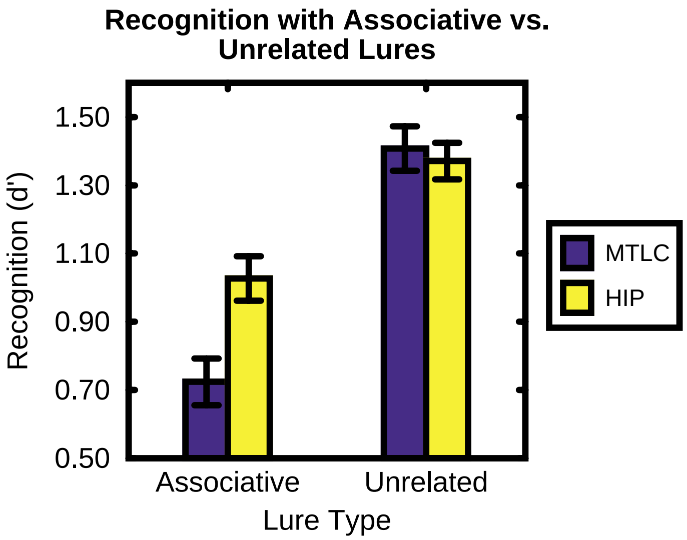
Using the combined model, we simulated an associative recognition experiment (using YN testing); we also ran a standard, "unrelated lures" recognition experiment for comparison.

Key results:

- Cortically-driven recognition is worse, relative to hippocampally-driven recognition, when related associative lures are used at test vs. when unrelated lures are used at test

=> this replicates the YN related lure deficit reported earlier

- The cortical network's ability to discriminate between studied items and associative lures was well above chance



=> this shows that cortex is sensitive to feature co-occurrence in addition to (individual) feature occurrence

Evidence consistent with this conclusion comes from recent studies showing good associative recognition performance in patients with focal hippocampal damage, sparing MTLC (Vargha-Khadem et al., 1997; Holdstock et al., submitted).

Summary and Conclusions

We used neural network models of the hippocampus and medial temporal neocortex (MTLC) to explore these structures' contributions to recognition memory.

We identified several manpulations that should differentially affect hippocampally-driven and MTLC-driven recognition: list strength, average between-item similarity, and the relatedness of lures to studied items.

These differences support the idea that the hippocampus and MTLC make qualitatively distinct contributions to recognition memory.

We cited a wide variety of empirical results in support of the model's predictions. Most strikingly, we cited new patient data confirming our model's prediction of a three-way interaction between MTLC vs. hippocampus, YN vs. FC testing, and related vs. unrelated lures.

The model provides provisional support for the three key assumptions of Yonelinas and Jacoby's dual-process framework. We plan to use the model to deliniate the conditions under which the assumptions break down.

More generally, this work brings together two streams of memory research -- precise, formal modeling of list-learning data, and cognitive neuroscience approaches to memory -- that, up to this point, have operated in parallel.

By establishing an explicit mapping between the brain structures involved in recognition, and the processes they support, this work makes it possible to:

- bring cognitive neuroscience constraints to bear on mechanistic models of recognition memory, and...

- use these models to predict cognitive neuroscience data (e.g., lesion effects, neuroimaging activations)

Future directions

The cortical model can be used to explain *priming* data -- we think that familiarity and priming reflect the same underlying mechanisms operating at different levels of the neocortical hierarchy.

The hippocampal model can be used to simulate *cued recall* (e.g., source memory data).

We can use the combined model to simulate:

- data on the *time course* of recollection and familiarity

- data from paradigms where recollection and familiarity are placed in *opposition* (e.g., Jacoby's process dissociation paradigm)

The combined model can be used to predict *neuroimaging activations* (simply by reading out the activation of different parts of the model in response to different test probes).

We should be able to precisely predict the effects of different kinds of *medial temporal lesions* by lesioning the corresponding part of the combined network.

We are using the same cortical-hippocampal model to account for the effects of hippocampal lesions on animal learning (O'Reilly & Rudy, 1999).

References

Aggleton, J.P., & Brown, M.W. (1999). Episodic memory, amnesia, and the hippocampal-anterior thalamic axis. *Behavioral & Brain Sciences, 22*, 425-489.

Curran, T. (in press). Brain potentials of recollection and familiarity. *Memory & Cognition*.

Holdstock, J.S., Mayes, A.R., Roberts, N., Cezayirli, E., Isaac, C.L., O'Reilly, R.C., & Norman, K.A. (submitted). Memory dissociations following human hippocampal damage.

Jacoby, L.L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory & Language, 30*, 513-541.

Jacoby, L.L., Yonelinas, A.P., & Jennings, J.M. (1997). The relation between conscious and unconscious (automatic) influences: A declaration of independence. In J. D. Cohen & J. W. Schooler (Eds.), *Scientific Approaches to Consciousness* (pp. 13-47). Mahwah, NJ: Erlbaum.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105*, 724-760.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102*, 419-57.

Norman, K. A. (submitted). Differential effects of list strength on recollection and familiarity.

O'Reilly, R.C., & Munakata, Y. (2000). *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*. Cambridge, MA: MIT Press.

O'Reilly, R.C., & Rudy, J.W. (1999). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. University of Colorado, Institute of Cognitive Science Technical Report 99-01.

Ratcliff, R., Clark, S.E., & Shiffrin, R.M. (1990). List-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 16*, 163-178.

Rolls, E. T., Baylis, G. C., Hasselmo, M. E., & Nalwa, V. (1989). The effect of learning on the face selective responses of neurons in the cortex in the superior temporal sulcus of the monkey. *Experimental Brain Research, 76*, 153-64.

Shiffrin, R., & Steyvers, M. (1997). A model for recognition memory: REM -- retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*, 145-166.

Vargha-Khadem, F., Gadian, D.G., Watkins, K.E., Connelly, A., Van Paesschen, W., & Mishkin, M. (1997). Differential effects of early hippocampal pathology on episodic and semantic memory. *Science, 277*, 376-80.

Yonelinas, A.P., Dobbins, I., Szymanski, M.D., Dhaliwal, H.S., & King, L. (1996). Signal-detection, threshold, and dual-process models of recognition memory: ROCs and conscious recollection. *Consciousness and Cognition, 5*, 418-441.

This research was supported by NIH NRSA Postdoctoral Fellowship 1 F32 MH12582-01, awarded to Kenneth A. Norman