Simulation and Explanation in Neurospychology and Beyond

Randall C. O'Reilly
Department of Psychology
University of Colorado at Boulder
Campus Box 345
Boulder, CO 80309-0345
oreilly@psych.colorado.edu

Martha J. Farah
Department of Psychology
University of Pennsylvania
3815 Walnut Street
Philadelphia, PA 19104-6196
mfarah@cattell.psych.upenn.edu

January 19, 1999

Submitted to: Cognitive Neuropsychology

Introduction

Like our colleagues Young and Burton (in press) (YB), we believe that good models explain a wide range of data, in ways that are motivated by independent theoretical considerations, and bad models explain a narrow range of data, by the *ad hoc* fitting of the model to the data, divorced from any more general theoretical considerations. Alas, YB's commentary demonstrates the difficulty of applying these seemingly straightforward principles to real models in a given research area. One needs an understanding of both empirical and computational issues before one can meaningfully judge "wide" versus "narrow" and "principled" versus "ad hoc." For example, accounting for a number of highly similar tasks should not be taken as evidence for "wide" explanatory scope, nor should explanations based on general computational principles be judged "ad hoc" because their independent motivation is not drawn from the realm of existing psychological models. We will argue that YB's preference for the IAC model (and localist models more generally) over our Farah, O'Reilly, and Vecera (1993) (FOV) model (and distributed models more generally) is based on a mistaken accounting of breadth of applicability and a neglect of fundamental computational principles, along with more prosaic errors such as a number of apparent mistakes in implementing simulations and a failure to note that several basic predictions of their model are disconfirmed by the available evidence. Underlying this broad pattern of failure to appreciate and attend to computational issues in modeling (from technical issues of implementation to theoretical issues of model predictions and neurobiological plausibility) is a fundamentally different view of the role of computation in psychological explanation. YB deny that features of the computation (such as the distributedness of the representations) are part of the model proper, and can play an explanatory role, instead relegating the computational aspects of psychological models to theory-irrelevant implementation.

We have organized our response into three parts that parallel YB's, addressing their three questions: 1) Which model gives the most complete account of covert recognition in prosopagnosia? 2) Which model has wider applicability to related phenomena in the literature on face recognition? 3) What are the relative merits of the different modeling styles? We begin by introducing and clarifying a central point of contention between the two modeling styles, the use of localist versus distributed representations. We return to this issue again in part 3.

In response to the first question, we point out that YB's model initially explained only a narrow range of data, and their new model explains two new phenomena only by using basic features of FOV, including distributed representations. Further, while their model now captures the two additional covert recognition phenomena, its predictions conflict with other more basic findings about prosopagnosia. FOV, on the other hand, provides a principled explanation of a disparate set of covert recognition tasks, including tasks that YB incorrectly state are beyond its scope, and also accounts naturally for several other features of prosopagnosia. In response to the second question, we show that their critique of our model is based on a number of mistaken beliefs about the capabilities of the FOV model and distributed representation more generally. In response to the third and most general question, we identify a basic difference of approach to modeling that appears to underlie the many other differences between YB's views and our own. Whereas YB regard computation as a tool for simulating already-articulated psychological theories, we view computation itself as potentially explanatory. We then present a sample of the overwhelming body of empirical and computational evidence supporting the reality of, and explanatory value of, distributed representation in human cognition. To YB's lament that distributed systems are more difficult to understand than local, we say "perhaps so," but while this is a relevant criterion for workers in the field of Human-Computer Interaction, it is not relevant for scientists selecting among theories of the natural world.

Explaining the Overt/Covert Dissociation

The strength of the FOV model, in our view, was that it explained the overt/covert dissociation in three fundamentally different tasks on the basis of some very general properties of distributed network computation. Thus a reasonably wide scope of data was explained without invoking any assumptions specifically for that purpose, but rather by showing that they are a natural consequence of independently motivated and commonly used assumptions concerning computation by neural networks. YB's characterization of FOV as a case of "the *ad hoc* development of models to account for specific phenomena" thus misses both the principled basis of the model's success (e.g., distributed representations were not invented by us for the purpose of explaining covert recognition) and the generality of its scope (three very different manifestations of covert recognition). As we will detail below, the model is also successful in simulating various additional types of tasks, including sequential associative priming, overt familiarity judgments, forced choice cued recognition, and provoked overt recognition, demonstrating an even wider explanatory scope. There is no possibility of *ad hoc* fitting in these cases, as we simulated these tasks only in response to YB's allegation that it could not be done, and produced successful simulations that depend on the same small set of principles as in our original model.

In contrast, the original IAC model explained behavior on just one general type of covert recognition task, which we originally termed "priming." Furthermore, as we will explain below, it did so by the ad hoc application of a decision criterion, external to the model itself and invoked only for overt tasks. Although the new version accounts for the two other tasks modeled by FOV, it does so with the help of two other features, for a 1:1 ratio of data to assumptions. With these features, shared with FOV (distributed representations and a learning mechanism), IAC can account for almost the same scope of covert recognition data as FOV. However, even with these features it makes a number of wrong predictions about prosopagnosia.

The FOV Account of Overt/Covert Dissociations

YB argue that the FOV model fails to account for three important aspects of overt/covert dissociations that the IAC model can account for: covert associative priming, overt familiarity judgments, and cued recognition. Here we show this to be wrong in all three cases. Further, we show that a fourth aspect, provoked recognition, which cannot be accounted for by the IAC model, can be simulated with the FOV model.

Sequential Associative Priming

When a face precedes a name by some short time interval, and the two are semantically related (e.g., both members of Britain's royal family), judgments about the name, such as a "famous" versus "not famous" judgment, can be made more quickly than with no face. This finding, called by YB "sequential associative priming," is also shown by some prosopagnosics, and is therefore a form of evidence for covert face recognition. In our 1993 article, we simulated a similar task involving simultaneous face and name presentations, and obtained associative priming as well as interference (delayed response to a name when the face is semantically dissimilar). Because the effects were so similar we grouped them together as one simulation of "priming".

As far as covert face recognition is concerned, there is no reason to distinguish between priming by an unrecognized face presented simultaneously with a name, and priming by an unrecognized face presented a few seconds before. In contrast, YB credit IAC with broad scope partly for its ability to simulate both effects, and allege that FOV cannot account for sequential associative priming.

Setting aside, momentarily, the question of whether FOV is really unable to simulate sequential associative priming, consider precisely what YB say FOV cannot do. They do not call our attention to a problem in priming a name judgment with a face, which FOV has already simulated. Nor do they report a failure to

obtain priming per se when the face precedes the name. Rather, they were unable to obtain any response to the name preceded by a face, and so could not determine whether a face would or would not prime a subsequent name. YB correctly point out that FOV's attractor states are so strong that subsequent inputs have little effect, making it impossible to simulate any task involving sequentially presented stimuli. This is a well-known problem for attractor networks, and would likely be a problem for our brains as well if not for such factors as the discrete spiking nature of real neurons as compared to the continuous, real-valued outputs of model units, and neuronal fatigue. Fatigue can easily be captured in the model by introducing activation decay after the network settles into an activation state. This is commonly done when networks are used to model sequential processes (e.g., Burgess, 1995; Dayan, 1998). Because we had not set out to simulate any sequential tasks, we did not originally incorporate decay. However, when FOV's activations are decayed after the presentation of the face stimulus, the subsequent presentation of the name stimulus was able to propagate through the network, and the presence or absence of sequential associative priming could be tested.

Using the original FOV model, we did not find a substantial priming effect, presumably because of the relatively tiny difference in amount of overlap between the semantic representations of people from the same category and different categories (only one unit). We therefore altered the patterns learned by the model to include semantic representations in which the members of the same category (e.g., "actors"), all had overlapping distributed representations constructed as random variations of a common prototype. With this greater within-category semantic overlap, the network exhibited significant sequential associative priming at levels of damage up to 75%, the same degree of damage at which the system performs at chance on an overt task. Note that changing the patterns in this way would not be expected to affect the qualitative pattern of results in any of the previously reported simulations. We confirmed this by replicating the results from the original model. Appendix 1 gives the modeling parameters and results.

Familiarity

Many of YB's criticisms of our simulation of the overt/covert dissociation hinge on tasks involving familiarity judgments. We intentionally avoided simulating such tasks, because they require the modeler to take a stance on the mechanistic basis for familiarity judgments. Although familiarity seems to be a simple concept, the ways in which subjects make familiarity decisions are anything but simple. Perusal of the psychological literature from memory research (e.g., Jacoby, 1991) to lexical decision (which concerns familiarity decisions about letter strings; e.g., Seidenberg, Waters, Sanders, & Langer, 1984) makes clear the variety of factors that come into play, including automatic processes of both a perceptual and conceptual nature, and strategic processes. Modelers have made various attempts to find reasonable computational interpretations of familiarity (Plaut, 1997; Becker, Moscovitch, , Behrmann, & Joordens, 1997; Mathis & Mozer, 1996; Metcalfe, Cottrell, & Mencl, 1992), but no consensus has emerged at this point. YB attempt a very easy solution, simply stipulating that familiarity is PIN activation. Rather than incorporate questionable assumptions into our model, we prefer to remain agnostic about the mechanisms of familiarity judgment. What we give up is the possibility of attempting to simulate some overt-covert dissociations, specifically those designed to include familiarity judgments.

Sensible people can disagree, and YB apparently place greater value than we do on simulating all variants of the overt-covert dissociation, as opposed to representative results from each type of task (relearning, priming, and speed of perception), even at the price of incorporating additional assumptions into a model. Therefore they attempted to simulate familiarity judgments with FOV, by assigning settling speed of visual or semantic units the interpretation of familiarity. They report that their simulations using this implementation of familiarity in FOV failed to capture the overt-covert dissociation. Given our reservations about the possibility of any simple implementation of familiarity, we were not surprised to learn of this failure.

We were therefore doubly surprised when we could not replicate their reported failure with FOV! Con-

trary to our own conservatism regarding the computational tractability of familiarity, and also contrary to the reported simulation results of YB, we easily found the overt-covert dissociation when speed of settling in semantic units was used as an overt familiarity measure in our model. Specifically, at 50% damage to the face hidden units, a level of damage at which the various covert measures simulated by us show positive evidence of covert recognition, the averaged results from 50 random samples of forced-choice familiarity decisions showed settling time for familiar faces was not significantly different than the settling time for unfamiliar faces; indeed it was nonsignificantly longer. Further, when we used a different measure of familiarity known as the *goodness* (aka negative *energy*) of the network's activation state, which has been used in several other models (Becker et al., 1997; Borowsky & Masson, 1996; Rueckl, 1995), we found that the advantage for trained ("familiar") faces over unfamiliar ones disappeared at only 25% damage. Thus, consistent with our reservations, the overt familiarity behavior of the model depends substantially on which familiarity measure is used. Nevertheless, two candidates for a familiarity measure both yielded the desired dissociation. Further details of these simulations are included in Appendix 2.

We do not know why YB did not obtain the same result using the settling time familiarity measure. Although they state that they have "attempted to capture loss of familiarity in forced-choice tests" they say nothing about their simulation attempts and at least part of their conclusion is based not on simulation but on the reasoning that "Whenever there is any residual effect of learning, the model will favor a known over an unknown pattern," which they support empirically by reference to our finding (FOV, 1993, Simulation 1) of faster settling in visual units for familiar patterns. This reasoning reveals a misunderstanding of the behavior of distributed interactive networks. The visual units may settle faster with familiar patterns after damage, but as long as they are settling into incorrect states, the inputs to the semantic units for familiar patterns may be as far from well-structured semantic attractor basins as the inputs for unfamiliar patterns.

In conclusion, we remain agnostic concerning the correlates of familiarity in neural networks, and therefore assign little weight to the success of the overt/covert simulations using either PIN activation or semantic settling time as a measure of familiarity. But to the extent that semantic settling time, or goodness, are reasonable candidates for familiarity in a neural network, the FOV model easily accounts for the dissociations in question. We cannot explain why YB did not obtain this result empirically, but note that their a priori reasoning was flawed concerning the impossibility of dissociating semantic settling time from covert measures.

Forced Choice Cued Recognition

Cued recognition is another form of covert recognition, in which prosopagnosics can make a correct forced choice decision between two names while viewing a face, even though they cannot overtly judge the face familiar or unfamiliar, or name the face. Contrary to the claim that FOV cannot be made to simulate this phenomenon, our model explains it very naturally, and we are glad for the opportunity to demonstrate FOV's success in another qualitatively different type of task.

The important thing to note is that the forced choice cued recognition paradigm provides a strong source of additional constraint on the settling process in the form of the name input to the semantic layer via an intact pathway. This name input is capable of producing the correct corresponding semantic representation by itself, whereas the face input via the damaged pathway is not capable of producing the correct semantic representation by itself. One must be careful not to confuse the behavior of the network with only the weak constraint provided by the damaged face input with that when both this weak constraint and the strong name input are provided. In the latter case (i.e., forced choice cued recognition), the weak additional constraints provided by the face input can have a measurable impact because the network is brought into a region of greater sensitivity to this input by virtue of the strong name input. In other words, the weak face input by itself produces something like a floor effect, and the additional name input brings the system off this floor so that the damaged face input can now have measurable effects.

This reasoning was confirmed by simulation using the FOV model. Using either semantic settling time or goodness as a measure of familiarity, we were able to simulate this cued recognition effect without any additional changes to the model, as described in Appendix 3. For example, at 75% damage to the face hidden units, where overt familiarity measures had long since failed, the system retained the ability to distinguish between correct and incorrect names for the faces.

Provoked recognition

Provoked recognition is another form of preserved face recognition in prosopagnosia, in which the subject ultimately experiences overt recognition. After seeing a number of faces from a single semantic category, such as actors, faces can be named and, reportedly, experienced as familiar. YB assert that neither IAC nor FOV can account for this finding, but in fact the phenomenon is compatible with a distributed constraint satisfaction architecture and can be simulated by FOV. The gist of the explanation is that repeated presentations of different faces with common semantic subpatterns will result in a build-up of residual activation primarily in that subpattern. This activation will sometimes provide the needed additional constraint to make up for the loss of constraints coming from damaged face representations to allow for successful semantic retrieval and naming.

In order to test this interpretation, we presented a set of face input patterns all from the same semantic category (with the same decay manipulation as used in semantic associative priming between each input), and recorded measures of naming and familiarity as before. We found that overt recognition was more likely to occur after viewing multiple faces from the same category, as measured by greater familiarity and, at some levels of damage, greater success in naming. Factors contributing to the size of the effect include the amount of semantic pattern overlap and the amount of decay used. Our face-semantic-name patterns were not optimally designed for this simulation, with a common semantic subpattern of only 2 units, and because the model was not originally set up to simulate sequential effects, a relatively large amount of decay was necessary to overcome the strong attractor dynamics of the network, which reduced the level of accumulated activation in those units. Despite these limitations, the effect is reliable. See Appendix 4 for simulation details and results.

The IAC Account of Overt and Covert Recognition in Prosopagnosia

So far we have seen that the FOV model is capable of explaining the full range of overt-covert dissociations discussed by YB, and that it does so in a natural way, without alterations designed solely for this purpose. We now turn to the IAC model, which differs both in failing to account for some of the key data on overt and covert recognition in prosopagnosia, and in relying on an ad hoc addition to the IAC model itself to account for the overt-covert dissociation.

Ad hoc Nature of the IAC Account

The original IAC model accounted for priming-based covert measures (associative priming and interference) by assuming that prosopagnosia uniformly attenuates weights from the face recognition units (FRUs) to the person identity nodes (PINs), and that overt task performance such as familiarity judgment requires that a *threshold* on the activation of the PINs be exceeded. The first time we read this, we assumed that this threshold was of the standard type used in neural network models, and could see how this might play a role in such a dissociation. But upon rereading, we realized that the explanatory work in this model is being done by a type of threshold that is unlike others discussed in the neural network literature — their threshold serves absolutely no computational purpose within the network, and its function is solely as an overt-covert-dissociation-maker.

What does it normally mean for a unit to have a threshold? Units in neural networks summate input activation and also pass on output activation to other units. In many networks, units only pass on activation if the summated activation exceeds a certain value — the unit's threshold. Real neurons also have thresholds

in this sense of the word. In the IAC model, however, activation is continuously cascaded between units during both overt and covert tasks. Thus, their overt-covert threshold is not about determining when enough activation has accumulated to be propagated onwards. This applies to all of the units in the IAC model, including the PINs, and indeed it is the continued output from the PINs, despite their attenuated inputs, that underlies the preservation of the covert priming tasks.

The PIN "threshold" that underlies the overt-covert dissociation in the IAC model is not part of the IAC model proper. Instead, it is a decision criterion applied to PIN activation levels only when they are used for overt familiarity judgments, and is external to the model, affecting none of the model's activations or weights. In the authors' own words, "Note that these threshold values are (of course) arbitrary. However, the exact thresholds chosen do not affect the processing of the model in any way. Activation is continually passed in a cascade fashion, and the threshold affects only the decision criterion" (Young & Burton, in press). *It is because, and only because, the overt task of familiarity judgment has been stipulated to involve a decision criterion, using this external-to-the-model, arbitrary "threshold" that the IAC model dissociates overt and covert recognition.* The essence of the IAC account of the overt-covert dissociation is this: "If one form of recognition is impaired after damage and another is spared, then hypothesize that an arbitrary criterion for minimal quality of processing is required just for the impaired ability and not for the spared one."

This account is unsatisfying in the same way that that accounts of overt-covert dissociations that feature a "consciousness box" are unsatisfying: While both account for the basic dissociation in a straightforward way, it is just a little too easy to explain a selective impairment in one type of task by postulating a special component of the mind (consciousness system or decision criterion) that happens to be required only for the impaired task, without any other, independent motivation for including that component in the model or involving it in just the impaired tasks. Indeed, although YB seem to regard the IAC account as an improvement over the earlier hypothesis that face recognition had been disconnected from a consciousness system, we do not. A "decision criterion" may sound more mechanistic than a "consciousness system," but we have already shown that it in fact plays no mechanistic role in the behavior of the model, whereas there is at least ample independent precedent for hypothesizing systems involved in conscious awareness.

Finally, we note that the ability of the IAC model to account for the two other tasks originally modeled by FOV depend on two additional assumptions, for a 1:1 ratio of model features to effects explained. A learning mechanism was added to model savings in relearning, and distributed face representations were added to account for familiarity effects in face matching, bringing the IAC account closer to FOV. Even with these features, however, the two models are not equally successful. In the four sections that follow, we review four of the IAC model's predictions about prosopagnosia that are clearly wrong.

IAC Predicts Intact Forced Choice Overt Recognition in Prosopagnosia

When psychologists suspect that performance in a task is limited by a decision criterion, that prohibits subthreshold knowledge from being expressed, they turn to a forced-choice paradigm. Instead of asking the subject "Is this an X, yes or no?" they show the subject and an X and a Y and ask "Which of these is an X?" Although the strength of the X-hood signal for the X might be below the criterion for deciding "Yes," it could still be greater for the X than for the Y. For this reason, accuracy in forced-choice tasks is sometimes called a criterion-free measure of subjects' ability (Green & Swets, 1966).

As we have already seen, the IAC model dissociates overt and covert recognition through the use of a criterion for PIN activation in an overt "yes/no" familiarity task. However, when we switch to this criterion-free forced-choice measure of overt performance, *the IAC model always produces perfect performance, even with "prosopagnosic" levels of damage (FRU-PIN link attenuation).* Evidence for this can be found in YB's Figure 7a, which shows that a familiar face will always cause higher PIN activation than an unfamiliar or less familiar face. Similarly, YB's Figure 4b shows that although the activation in name units might be too low

to exceed a decision criterion after FRU-PIN link attenuation, the unit for the correct name will always be more active than the units for incorrect names, predicting accurate forced choice among names. Of course, we know that the overt recognition impairment of prosopagnosia is just as evident on forced-choice tests as on "yes/no" and naming tests. Figure 3 in our original 1993 paper shows that FOV reaches chance levels of performance on a forced choice task above 50% damage while continuing to manifest covert recognition by a number of measures.

IAC Predicts Prosopagnosia is All-or-none

Neuropsychological disorders can be mild or severe, and may change their level of severity over time. After an acute injury, the disorder may be severe and then gradually recover, either partially or completely. In degenerative conditions, the reverse may be seen. There is no neuropsychological impairment that is seen only in full-blown form or not at all. In particular, prosopagnosia can exist in mild, moderate or severe forms. Yet the IAC model predicts that patients are either normal at overt face recognition or totally unable to recognize any faces overtly. This problem results directly from the way the model accounts for covert recognition, namely the combination of a threshold on local PINs and uniform weight reduction between face recognition units (FRUs) and PINs. As the weights are attenuated, overt performance will remain unchanged until the familiarity threshold is reached. At that point performance will drop to chance levels, and remain there.

One might try to fix this problem by making the weight reductions nonuniform. For example, the most realistic way of implementing damage in a network would be to eliminate some connections altogether while leaving others intact, allowing levels of overt recognition performance to fall anywhere between perfect performance and chance depending on the proportion of connections eliminated. Unfortunately, this implementation of damage eliminates the overt-covert dissociation: Some faces will be recognized, both overtly and covertly because their FRU-PIN connections are intact, and others will not be recognized either overtly or covertly, because their FRU-PIN connections have been severed. In light of this problem, one might aim for intermediate overt performance by varying the degree of attenuation of FRU-PIN connections without eliminating connections. This would have the desirable result of overt performance measures intermediate between perfect performance and chance, with the possibility of covert recognition for faces not overtly recognized. Unfortunately, it has the undesirable result of predicting perfect test-retest reliability, that is, certain faces always recognized and all other faces never recognized. Weak item effects may be seen with some prosopagnosics, but it is not the case that certain faces are reliably recognized, across different depictions, whereas others are never recognized. A final solution is to combat the perfect consistency of the model by directly adding variability to the activation values of the units. Although this could produce intermediate overt performance without strong item effects, it would be yet another ad hoc addition to the model.

IAC Predicts Prosopagnosia Affects only Familiar Faces

Although the literature contains claims of selective impairment of familiar face processing in prosopagnosia, whenever the perception of unfamiliar faces has been carefully tested it has been found to be impaired (see Farah, 1990; Shuttleworth, Syring, & Allen, 1982, for reviews). Young, Newcombe, de Haan, Small, and Hay (1993) have shown that apparent dissociations between the processing of familiar and unfamiliar faces disappear when time to perform perceptual tests with unfamiliar faces is taken into account; patients may achieve a good accuracy score by abnormally slow and slavish checking of facial features. Indeed, cases PH, in whom covert face recognition has been demonstrated by preserved familiarity effects in face matching, performs simple perceptual face matching poorly (16% errors) and slowly (almost 3 seconds on average) even when the faces are unfamiliar (de Haan, Young, & Newcombe, 1987).

In contrast, the IAC model is based on the assumption that the impairment in prosopagnosia lies downstream from the perceptual representation of faces, in a part of the system that exists only for familiar faces,

namely connections between the perceptual FRUs and the PINs. The IAC model could be defended by hypothesizing that, for reasons of anatomical proximity, visual face representations are also likely to be damaged in cases of prosopagnosia, and have so far invariably been damaged. The FOV model has the advantage, however, of not requiring coincidental damage to two parts of the system; it is based on the assumption that perception of faces, familiar and unfamiliar, is impaired in prosopagnosia.

IACL Predicts Prosopagnosia is Temporary

The addition of a learning mechanism to the IAC model, resulting in the IACL model, creates another problem: it commits the modelers to the prediction that prosopagnosia is temporary, in that it can be entirely overcome by relearning. Given the way damage and relearning are simulated in IACL, there is nothing that requires the relearning to stop short of perfect performance. Indeed, comparing the results shown in their Figures 4a, b and c, one can see that after 5 trials of relearning, the network has completely recovered to an unlesioned level of performance. This would predict that prosopagnosic patients could recover all of their lost knowledge by simply studying all the faces they once knew for some (apparently relatively short) period of time! In contrast, relearning in FOV has a low asymptote, because a reduced number of units and weights are available to store the new knowledge — the network, like prosopagnosics, has actually suffered irreparable damage.

IAC's Incorrect Predictions Follow from Theory-Relevant Features

Of course, for every scientific model, some features are theory-relevant and some are not. We have highlighted several ways in which the predictions of the IAC model fail to accord with reality. An important question to ask is whether the IAC model's incorrect predictions result from theory relevant or theory-irrelevant features. In all cases, the failures derive directly from theory-relevant features, and directly or indirectly from the use of local representations.

In both models, covert recognition is the result of a partially functioning system. With FOV's distributed representations, the "partiality" of the system's knowledge of faces consists of a subset of the weights that originally embodied knowledge of the faces' appearance. There is no equivalent way of damaging face representations with IAC's local representations, and so the partiality of functioning instead results from attenuated connection strengths between FRUs and PINs. This difference in the way partial functioning can take place in distributed and local systems accounts for all of the problems that the IAC model encounters in simulating prosopagnosia. The attenuation of FRU-PIN connections cannot account for impaired overt recognition without the imposition of a decision criterion external to the model, but this leaves the model unable to account for impaired overt recognition in criterion-free tasks. The choice between all-or-none prosopagnosia and strong item effects is forced upon the IAC model by its use of local representations, in conjunction with the criterion needed to create the overt-covert dissociation. Either weights are uniformly attenuated, giving rise to the all-or-none problem, or they are nonuniformly attenuated, giving rise to the perfect test-retest problem. There is no natural way to obtain a gradient of performance with varied amounts of damage other than having one specific face at a time drop from the "always recognized" to the "never recognized" category without building in variability explicitly for this purpose. In contrast, with distributed representations, each of the units and weights participate in the representation of many faces, and damage to each unit or weight therefore impacts on many faces. And because each face is represented by many units and weights, damage to each unit or weight has only a moderate effect on recognition of that face. Increasing damage therefore results in a gradient of performance for all faces, and because any particular lesion may by chance affect more of the units and weights involved in one face's representation than another's, there may be weak item effects.

The different locations of the lesions in the IAC and FOV models, and their consequent predictions concerning unfamiliar face processing in prosopagnosia, can also be traced to the difficulty of implementing partial or graded performance in systems of local representation. For the reasons just stated, distributed face

Effect	**FOV**	Comments	**IAC**	Comments
10AFC Overt at Chance	Yes		No	Always 100% correct
Graded Levels of Prosopagnosia	Yes		No	All-or-nothing
Unfamiliar Faces Impaired	Yes		No	Only familiar have PINs
Permanent Effects of Damage	Yes		No	Complete recovery
Savings in Relearning	Yes		No	Complete relearning
Speed of Settling (Perception)	Yes		Yes	Uses distributed reps
Semantic Priming (& Interference)	Yes		Yes	
Sequential Associative Priming	Yes	Requires decay	Yes	
Forced-choice Cued Recognition	Yes		Yes	
Provoked Overt Recognition	Yes	Requires decay	No	

Table 1: Corrected version of Young & Burton's Table 1, showing accurate comparison between the two models on basic properties of prosopagnosia, and the set of covert measures considered.

representation naturally accommodates partial processing of all faces after damage. Because their model does not include distributed face representations, the authors of the IAC model were forced to interpret partial processing in terms of weakened connections between intact face representations and post-perceptual representations downstream.

IACL's prediction that prosopagnosia should be only transitory also follows from the use of local representations, whereas FOV's prediction of a low asymptote for relearning after damage follows from the use of distributed representation. Knowledge in FOV is represented in a distributed manner across a large number of weights. When the FOV model relearns, it must compensate for the permanently missing units by reusing the remaining weights and units. Thus performance is constrained to remain permanently impaired because of the reduced number of units and connections. It is also worth noting that learning in FOV is an integral part of the model, because distributed representations cannot be hand-wired as is possible with local representations, and was not simply added to the model to account for particular data. In IACL, the simulation of damage by attenuating connections and relearning by strengthening them again fails to put any constraints on the amount of relearning that can be achieved.

Young & Burton's Table 1, corrected

Our goal in writing this article is to clarify certain aspects of the behavior of distributed interactive neural networks, and to discuss the relevance of this behavior to psychological explanation in the case of covert face recognition. The detailed accounting of which model can simulate which variant of which task is of less interest. Nevertheless, the two concerns cannot be entirely divorced from one another, and so we wish to set the record straight on the successes and failures of the models that YB contrast in their Table 1. Thus, we provide a corrected Table 1, revised to include all of the paradigms we simulated in our original paper, together with the new paradigms considered here.

Accounting for the Phenomena of Normal Face Recognition

YB find it problematic that a model aimed at explaining the covert/overt dissociation does not explain other phenomena related to face recognition. But the motivation of FOV was to demonstrate that overt/covert dissociations could be explained simply in terms of some basic and general properties of neural information processing, not to explain face recognition more generally. Do we think models deserve more credit for explaining a wider range of phenomena? Yes, of course. Do we think that models are suspect if they do not

explain phenomena outside the realm originally intended? Not at all, particularly if there is nothing in the model that would conflict with an attempt to broaden its scope.

The FOV model is perfectly compatible with the broader range of face perception phenomena reviewed by YB, their claims to the contrary. Just as they and their colleagues were able to simulate additional findings in person perception by adding features to the original IAC model, such as separate input and output name representations, and a direct route from name inputs to name outputs, the same is true of FOV. Neither model accounted for the full range of data reviewed by YB in its original form. Both models can do so with appropriate additions, and FOV requires no more additions than IAC. Indeed, the core features of the FOV model — distributed representation, learning, and interactivity — provide natural explanations of many of the phenomena in normal person perception that YB review. Here we consider each of the effects reviewed by YB in turn.

Repetition Priming

Repetition priming is domain-specific, in the sense that repeated judgments of an individual's face show priming whereas a judgment of a name followed by a judgment of the same individual's face does not show priming. An obvious explanation for this aspect of repetition priming is that the name representations engaged in the act of reading a name are distinct from the name representations engaged in the act of producing a name, for example when a face is named. This distinction between input and output lexicons, which is supported by considerable evidence in cognitive psychology and neuropsychology, implies that different name representations are activated when reading a name and when evoking a name from a face, and hence explains the absence of priming between these two types of task. Separate input and output lexicons were incorporated into the IAC model specifically in order to account for the domain-specificity of repetition priming, and as YB point out, the same could be done for FOV. We therefore fail to see what bearing repetition priming, and its likely explanation in terms of separate input and output lexicons, has on the issue of localist versus distributed representation in general, or on the choice between the IAC and FOV models in particular.

To put it another way, there is only one difference between the models that is relevant to their different abilities to simulate the domain-specificity of repetition priming: the IAC model to which separate lexicons were added *has* separate lexicons, whereas the FOV model does not. This is not a difference of "modeling style," but simply of what has been modeled. We trust that most readers did not interpret our use of a single lexicon as a theoretical statement that the same representations underlie the hearing, reading and speaking of names, but instead recognized that the distinction between input and output lexicons was simply not relevant to any of the tasks that the FOV model was intended to simulate.

A related point concerns the inexorable way in which FOV, and most neural networks generally, tend to complete patterns, resulting in all faces being automatically named. For the tasks we simulated, it was not necessary to incorporate an attentional mechanism that gates the flow of activation in the network depending on task demands, but such mechanisms are certainly available (e.g., Cohen, Dunbar, & McClelland, 1990), and provide a principled basis for the lack of transfer between face and name priming. However, given that we know that the presentation of a face will also automatically spread to the name units in the IACL model (as is necessary for naming, and priming effects), it is entirely unclear why their model does not predict cross-domain repetition priming as well, since they state that their Hebbian learning mechanism was applied to all the units in the network. Either they have some additional mechanism which suppressed the name unit activation, or they didn't in fact apply Hebbian learning to the entire network, or we are missing something about their model. The further discussion about "staged" processing and independent priming of different pathways in the IACL model, while at the same time advocating an ubiquitous link update proposal, seemed equally contradictory and confusing. We would appreciate a clarification of these points.

Asymmetrical Interference

The issue of asymmetry of interference effects is also an issue of what is modeled, rather than how. Let us set aside models altogether for a moment, and ask: Why is there less interference from a distractor face on reading a name than there is interference from a distractor name on naming a face? Because reading is different from naming! Specifically, it is possible to pronounce a letter string without knowing what it means, whereas one cannot name a face or object without having recognized it. Now let us return to models. This difference between reading and naming can be implemented in either the IAC model or the FOV model simply by inserting a direct route between name inputs and spoken name outputs. With the addition of this direct route, naming will be faster and less susceptible to interference from activity in semantic representations.

Thus, as before, there is only one difference between the models that is relevant to their different abilities to simulate interference asymmetry: the IAC model, to which was added a direct route from word inputs to word outputs, *has* a direct route from word inputs to word outputs, whereas the FOV model does not. There is nothing in the FOV model that would prevent it from being augmented in the same way as IAC was augmented, and hence the models are on equal footing. Indeed, the prototypical "radical" distributed connectionist model cited by YB, the Seidenberg and McClelland (1989) word reading model (and its descendants; Plaut, McClelland, Seidenberg, & Patterson, 1996) clearly advocates both a direct and indirect (via semantics) route between word input and output.

Time to Make Decisions

YB report that, in general, familiarity decisions to faces are made faster than semantic decisions (e.g., actor *vs* politician), which are in turn made faster than name decisions (e.g. John *vs* Richard). In contrast, when shown a name, name decisions are fastest, followed by familiarity decisions, followed by semantic decisions. YB allow that both the IAC and the FOV models can account for the pattern of decision latencies to faces quite naturally. It is the pattern of decision latencies to names that they claim poses a problem for FOV.

As with the finding of asymmetric interference, the finding that name decisions can be made fastest on names is a result simply of the direct route between name inputs and name outputs. The question "Is John Guilgud's first name John or Richard?" is rather like the question "Who is buried in Grant's Tomb?". While we do not believe the topic is entirely trivial, we also do not believe it has any bearing on the choice between localist and distributed models. Again, the only difference between the models that is relevant to their different abilities to simulate the pattern of decision latencies on names is that the IAC model, to which was added a direct route from word inputs to word outputs, *has* a direct route from word inputs to word outputs, whereas the FOV model does not.

Patterns of Error

Like decision latencies, error patterns reveal a ordering among types or stages of processing. It is possible to find someone familiar but fail to retrieve semantic or name information, or to find someone is familiar and retrieve semantic information about them but fail to retrieve a name, but it is never the case that a name can be retrieved without semantic information. Here too, YB allow that FOV and IAC can both accommodate this pattern naturally. But they claim that the two models diverge in their ability to explain error patterns in the case of a particular neurological patient, ME.

ME was unable to retrieve semantic information about familiar people from either a name or a face. Despite this, she could match names with faces with a high level of accuracy. YB take this as evidence for PINs, interposed between face and name representations, and separate from the semantic units which are

the hypothesized locus of damage in this patient. They conclude that a model like FOV, in which names and faces are associated solely by way of semantic representations, could not account for this effect. To quote:

> There is an architectural problem with FOV that leads to its failure to capture these data... If the semantics units [after damage] may be used for matching [faces to names], then they must be able to be used to pass information between the two modalities... This gives rise to a pattern of error which is never observed in humans, the 'name only' error. (YB, p. 60)

Despite their reference to FOV's "failure to capture these data," YB apparently did not attempt a simulation. Had they done so, they would have found that the model, exactly as described in our 1993 paper, does capture the data. Using semantic settling time as a type of familiarity measure once again, we implemented the matching task as a judgment of the *familiarity of the pairing* between face and name. Note that this is essentially the same as the cued-recognition task described previously, and the explanation for FOV's successful performance is exactly the same as well — multiple input constraints (from name and face inputs) can produce significantly better performance than one input constraint (either name or face alone). Thus, when semantic units are damaged, even to levels (e.g., 75%) at which the naming of a face is effectively at chance, the settling times for correct and incorrect name-face pairings remained discriminably different. Details of the simulation are given in Appendix 5.

Distinctiveness

Distinctiveness and similarity play a ubiquitous role in all aspects of perception and cognition, face processing included. As YB explain, a distinctive face is one with less overall similarity to other known faces. The distributed representations of FOV provide a natural way of understanding the many similarity-based phenomena of face processing, including associative priming, the classification of error types based on target-error similarity (e.g., semantic versus visual errors) and distinctiveness effects. This is because distributed representation provides a straightforward metric of similarity, and even more important, a mechanism for explaining its effects.

In distributed representations, greater overlap in active units corresponds to greater similarity. The greater confusability of similar representations arises from the fact that they are actually partially identical. Therefore, fewer units' activation values need be mistaken in order for one representation to be confused for another by the system when the two are similar. In contrast, pure localist representations are incapable of accounting for similarity-based effects without additional mechanisms added for this specific purpose. Indeed, YB's adaptation of the IAC model accounts for distinctiveness by using distributed representations of faces in a pool of feature units. We do not know what YB had in mind when they wrote "Using distributed representations, it is hard to capture the notion of a distinctive face," but surely they have misspoken.

Why were YB unable to obtain distinctiveness effects with the FOV model? Distinctiveness effects will be measured in terms of processing differences between face patterns that have small and large average distances from the other faces in the face similarity space, assuming distances among the patterns in semantic and name similarity spaces are roughly equivalent. This requires the existence of patterns with large and small distances. With the small set of patterns used in our simulations, and the small similarity space generated by having a total of only 16 face input units, the odds are strongly against finding subsets of face patterns whose average distances from other face patterns differ appreciably while avoiding offsetting confounds in semantic and name portions of the pattern. Indeed, when we calculated the average hamming distances of the face patterns in our simulation, they were tightly clustered around an overall average of 13, with the entire range only spanning from 12 to 14.4. With so little difference between the distinctiveness of the least and most distinctive faces in our patterns, and no control over the similarity relations among the non-face portions of the patterns, there is no wonder that YB did not find distinctiveness effects. Just as

Effect	**FOV**	Comments	**IAC**	Comments
Repetition Priming	Yes	w/sep lexicons	Yes	w/sep lexicons
Asymmetric Interference	Yes	w/sep lexicons	Yes	w/sep lexicons
Time to Make Decisions	Yes	w/sep lexicons	Yes	w/sep lexicons
Patterns of Error	Yes	Including ME	Yes	
Distinctiveness	Yes		Yes	Uses distributed reps

Table 2: Corrected version of Young & Burton's Table 2, showing comparison between the two models on ability to account for other effects in face and person recognition. Note that most of the effects are due to the addition of separate lexical representations and pathways.

the IAC model's patterns were designed to fall into groups with high and low distinctiveness for purposes of simulating distinctiveness effects, so FOV would need an appropriately designed set of patterns for that purpose.

The Issue: Explanatory Scope

YB raised the five topics just reviewed (repetition priming, asymmetry of interference effects, time to respond, error patterns, and distinctiveness effects) as evidence bearing on the explanatory scopes of the IAC and FOV models in particular, and of localist and distributed computational architectures more generally. It is to this issue, as opposed to the ins and outs of particular simulations, that we now return.

All parties presumably agree on the following criteria. If one model or modeling style can more easily accommodate the five assorted findings reviewed above, that would count in its favor. Furthermore, if one model or modeling style is actually constrained to predict some of the findings, that would count even more strongly in its favor. Finally, the five findings selected by YB constitute only a small subset of the potential explanatory scope of models of covert and overt face recognition, and the ability of the models and modeling styles to account for or predict other findings is also relevant.

So how do the models fare by these criteria? YB state that "We have shown that the FOV model is not generalizable to other phenomena found in prosopagnosia or for normal face recognition, whereas IAC can readily do this for a number of phenomena" (p. 75-76). Our accounting is different. In each of the five cases of normal person perception that YB selected, FOV is on equal or better footing than IAC. In three of these cases (repetition priming, asymmetry of interference effects, time to respond) either model can be made to account for the effects in question by specific additions, specifically separate name inputs and outputs and a direct route between them. These features are were not in the FOV model or the original IAC model, but could be added to either. They have no particular compatibility or incompatibility with local versus distributed representations. For a fourth finding, concerning error patterns, YB stated that FOV could not account for the key findings of case ME, but in fact the model simulated these findings without any additions or changes. In the case of the fifth finding reviewed by YB, distinctiveness effects, we have a case of one modeling style being actually constrained to predict the effect, as opposed to merely being able to accommodate it: distributed representations cannot help but generate distinctiveness effects, and indeed the IAC model accounts for these effects by adding a pool of distributed face representations. Thus, for the five findings that YB selected for comparing the models, FOV and distributed representations are overall slightly more successful. We have therefore provided an updated Table 2, corresponding to YB's Table 2.

What about findings beyond the five selected by YB? As already mentioned, effects involving similarity are explained in a straightforward way by distributed representations. Such effects include semantic priming, the greater tendency to confuse persons with similar faces, semantic information or names, the interpretation of error types, and the effects of visual, semantic or name distinctiveness. As explained in the first section, the existence of degrees of prosopagnosia, the permanence of prosopagnosia, and the impairment

of unfamiliar face processing in prosopagnosia are all explained naturally by FOV, based on the properties of distributed interactive computation, and are not accommodated by IAC. Finally, the FOV explanation of covert face recognition has explanatory scope along a different dimension as well: Its basic principles can explain other forms of performance dissociation between different tasks that ostensibly test the same knowledge. Mayall and Humphreys (1996) adapted the basic architecture to provide a novel account of preserved reading in pure alexia, and Munakata, McClelland, Johnson, and Siegler (1997) used the same underlying principles to explain why infants can manifest certain perceptual knowledge when tested in one way but not another.

Two Approaches to Computational Modeling

YB describe the function of the IAC model, relative to non-computational models in psychology, as follows:

> The difference between an implemented model like IAC and an unimplemented functional model such as Bruce and Young (1986) is only in the level of detail which each allows. The implemented model requires a greater level of specificity, and allows interactive exploration in a way which is not possible with box and arrow models (ms. p. 63).

In other words, YB approach the computational model as a powerful book-keeping device, forcing psychologists to be specific in their theorizing and enabling them to derive predictions from their functional models that exceed human working memory capacity.

We agree that this is one useful role that computation can play in psychology. But our approach is based on the idea that connectionist modeling has the potential to offer much more. We believe that in some cases, the correct explanation of a psychological phenomenon will be based on *properties of the computational system itself.* By this, we mean to exclude aspects of models that are shared with unimplemented cognitive psychology models, such as the direct route between word inputs and outputs in the elaborated IAC model account of aspects of repetition priming, asymmetrical interference, and time to make decisions. Rather, we are referring to properties to which the modeler is typically forced to make a commitment only when implementing the model, and whose reason for being is the actual information processing of the model, possibly also constrained (as in the case of many connectionist models) by a desire for neurobiological plausibility. Examples of such properties are distributed representations and the learning mechanisms that shape them.

For purposes of book-keeping there is little difference between local and distributed representation other than what YB call "transparency." But when the system being modeled is computational, as is the case with the human cognitive system, then the choice of a model's computational architecture has theory-relevant consequences for the behavior of the model. Some key behaviors that distributed representations commit us to are discussed in the following section. These are among the very behaviors that we, as psychologists, want to explain. We therefore place much less emphasis than YB on the "human-computer interaction" factor of transparency, and much more on two other considerations, discussed in the two sections that follow: the explanatory power of distributed representation and related features of connectionist architecture, and correspondences between these model features and real brain function.

Functional Differences between Distributed and Localist Representation

Distributed representations have many computational advantages over localist representations, which have either direct or indirect implications for psychology. These advantages result from the use of overlapping subsets of units to represent different entities. Note that this is the essence of what makes a representa-

tion distributed. The interpretability of individual units does not affect these properties, and is not a criterion for distinguishing distributed from local representation. Whereas in local representations individual units by definition have an interpretation, in distributed representations they may or may not be interpretable in simple or intuitive ways. Most of the units in our model cannot, but the actor and politician units are exceptions, and consititue parts of the distributed semantic representation. For an example of a distributed semantic representation in which each unit has an English-language interpretation, see Hinton and Shallice's (1991) model of deep dyslexia.

We begin our brief review (see Hinton, McClelland, & Rumelhart, 1986, for a fuller discussion) with the computational advantages of distributed representations that are of psychological relevance simply insofar as the brain and mind are likely to be optimized:

Efficiency: Fewer total units are required to represent a given number of input patterns if the representation is shared across units (otherwise one unit per pattern is required). For example, all of the colors can be represented by as few as three distributed units (e.g., red, green and blue), whereas localist representations would require as many units as color distinctions being made (i.e., 100's). This is particularly important as the domain becomes *combinatorial* in some feature space, since the number of possible different combinations (and therefore localist units) quickly approaches astronomical values.

Robustness: Having multiple units participating in each representation makes it more robust against damage, since there is some redundancy. For example, if one lost the "green" unit in a distributed representation, you would still be able to tell that red, yellow, and orange are similar . If you lose a localist unit, whatever knowledge was associated with that unit is completely gone.

Accuracy: In representing continuous dimensions, distributed (*coarse coded*) representations are much more accurate than the equivalent number of localist representations, because there is a lot of information contained in the relative activities of the set of nearby units, whereas localist units can only n different values (at an accuracy of $1/n$) for n units.

The following two ubiquitous psychological properties are also properties of distributed representation. Models incorporating distributed representation need not "build in" these properties. Rather, the properties are almost inescapable in systems of distributed representation:

Similarity: Distributed representations provide a natural means of encoding the similarity relationships among different patterns as a function of the number of units in common (*pattern overlap*). For example, orange can be represented as similar to red and yellow by virtue of sharing similar values of "red" and "green" components of a distributed representation. In contrast, each localist unit is an island, and the network would not naturally convey the fact that the orange unit is somehow related to the red unit.

Generalization: A network with distributed representations can often respond appropriately to novel input patterns by using appropriate (novel) combinations of hidden units. This is impossible for localist networks, which would require the use of an entirely new unit. For example, if we knew that red and yellow were associated with flame (and not to be touched), but for some reason had never seen an orange flame, we would still be able to generalize our response to this case based on the overlap of the color units involved. In contrast, one has to learn each association anew with localist units, which could be rather painful!

Neural Plausibility: Theory Relevant versus Theory-irrelevant Features

YB state that both FOV and IAC are both models of cognitive functions as opposed to neural structures (ms. p. 63), but this is not quite right. FOV, like other connectionist models, has some properties in common with real neural networks, and can therefore be considered a model of the brain. Of course, connectionist models also fail to incorporate much of what we know about the brain. In this sense, they are like any model in science, with both theory-relevant and theory-irrelevant properties. The theory-relevant properties are those properties of real neural computation, that have the potential for explaining certain psychological phenomena in terms of neural function. Among the theory-relevant properties of connectionist networks are: the large number of inputs to and outputs from each unit, the modifiability of connections between units, summation rules, bounded activations, thresholds, and the use of distributed representation.

Evidence for the brain's use of distributed representation comes from such indirect observations as the relatively global effects of damage to a given functional area (i.e., recognition of all faces is affected in prosopagnosia, not just some), and the "graceful" nature of the degradation (i.e., tissue loss may cause only mild or moderate impairments in face recognition), as well as more direct single cell recording experiments. A particularly relevant example of single cell evidence concerns the face cells of monkey temporal cortex. At first glance these might appear to be just the "grandmother cells" of localist caricature — cells that respond only when a single person, such as one's grandmother, comes into view. However, each face cell responds to a range of faces, and each face evokes activity in a sizeable fraction of the face cell population, consistent with distributed rather than localist representation (Desimone & Ungerleider, 1989; Young & Yamane, 1992). In a number of other domains of processing, analyses of cells' breadth of tuning and proportion of cells active suggests that distributed representation is ubiquitous in the brain. In motor systems, positional information related to the target for arm movements is coded in a distributed fashion (Georgopoulos, 1990), as is the direction for eye movements to a visual target (Sparks & Mays, 1990). Within the dorsal visual system, location is also coded in a distributed manner (Andersen & Zipser, 1988).

The neural plausibility of connectionist networks is crucial to the larger explanatory role that we give them, relative to the "book-keeping" role advocated by YB. If connectionist networks had nothing in common with real neural networks, but were simply a means to implement cognitive models (e.g., comparable to a programming language like LISP), then we would tend to agree with YB's statement that "very little of the explanatory power of a model should reside in the technical aspects of the architecture" (ms. p. 71). However, to the extent that the model architecture has significant commonalities with brain architecture, why would one eschew architecture-based explanations of cognitive functions of the brain? If a cognitive phenomenon that is ultimately based on underlying brain function can be explained in terms of other independently established principles of brain function, that would seem to be a reassuring sign for both the particular explanation in question and the principles invoked in that explanation.

Transparency: An Issue for HCI, not Science

YB's preference for local over distributed representation is based on differences in "transparency," or ease of interpretation (and implementation). They point out that additional measurement techniques are often required to interpret the outputs of networks with distributed representations, and they also criticize the use of bias weights for decreasing models' transparency.

We are reminded of the old story about the drunk looking for his keys under the street lamp, where the light is good, even though he dropped them by his door. Yes, it's easier to work with local representations, but if the goal of our work is a theory of human cognition, and human cognition is better explained using distributed representations, then ease of use seems irrelevant. There is indeed good reason to believe in the psychological reality of distributed representation. The physiological evidence reviewed in the previous section shows that the relation between neurons representing categories of stimulus or response and the

categories themselves is a many-many, distributed form of representation, not a one-to-one, localist form of representation. For those who think it hasty to conclude that properties of brain computation are also properties of cognitive computation, there is also the fact that distributed, but not local, representations display such fundamental psychological phenomena as similarity, generalization, and pattern completion. Finally, the success of specific distributed models such as FOV, over localist equivalents such as IAC, in accounting for specific psychological phenomena (see Tables 1 and 2) argues for the reality of distributed representation. In sum, given the empirical evidence that distributed representation plays a role in human cognition, as scientists we must rise to the challenge of building models that reflect this fact, whether or not the resulting models are easy to understand.

Learning

Learning can be used in connectionist models in two very different ways. It can play a theory-irrelevant technical role, as a way of algorithmically obtaining a network with weights that can perform the tasks of theoretical interest. It can also play a theory-relevant role, representing human learning. Let us first address some mistaken ideas about the technical aspects of learning contained in YB's critique.

YB's assertion that learning is a desirable component of a connectionist model only if it is intended to capture something about human learning reveals a naïvté concerning the practicalities of distributed models. Hand-wiring is impossible for all but the most trivially simple distributed systems, and learning is therefore essential to all distributed modeling, whether the theoretical scope of the model encompasses learning or is confined to a particular end state. Indeed, not only must learning be used in the set up of most connectionist models of the cognitive end state, it provides insights into that end state. Even if we view learning algorithms as purely technical tools for creating representations that enable a network to perform a task, they can explain why representations are as they are in terms of critical aspects of the task, the inputs, and the network architecture (e.g., the role of statistical features of words such as frequency and regularity in the representations underlying reading, Plaut et al., 1996).

The relearning simulations in FOV are intended to model learning per se, using the Contrastive Hebbian Learning (CHL) algorithm. YB criticize our use of CHL on the grounds that its learning is nonmonotonic, and therefore more difficult for the modeler to use. This is an HCI issue which, we reiterate, is irrelevant to choosing among cognitive theories. They also raise the more relevant question of whether human learning also has nonmonotonicities. Although YB remain agnostic on this question, and it has rarely been addressed directly by cognitive psychologists, there is evidence in at least one case of such a pattern: Marcus, Pinker, Ullman, Hollander, Rosen, and Xu (1992) have shown that nonmonotonicity is the rule in the acquisition of past-tense morphology at both the fine-grained and larger time scales.

YB's final criticism of learning in FOV applies to all algorithms for learning in distributed systems, not just CHL, and concerns the use of what they call batch learning. Because the same set of weights is changed every time a new item is learned in a distributed system, care must be taken to avoid unlearning previous items when a new item is learned. This is accomplished by cycling through all of the items to be learned (the entire "batch," in YB's terms) multiple times and using relatively small weight changes each time. YB point out that people do not do all of their learning in life in one batch on one occasion, a fact we do not dispute. However, current theorizing on memory and the brain holds that item learning is not confined to occasions on which the item is encountered. Rather, the process of memory consolidation involves an internal replaying of previously encountered items, interleaved with current items, to allow the gradual adjustment of shared weights (McClelland, McNaughton, & O'Reilly, 1995; Rolls, 1989). Thus, there is no incompatibility between learning in FOV or other distributed systems and the observation that people can learn new items at any point in life.

Finally, although YB neglected to raise the perennial critique of error-driven learning mechanisms as being biologically implausible compared to Hebbian learning, we note that many of these issues have been

recently addressed in an analysis that shows the close relationship between CHL and error backpropagation (O'Reilly, 1996). This analysis shows that whereas backpropagation requires the biologically implausible propagation of an error signal, which is a difference between two terms, CHL instead propagates the two terms separately as activation signals, and then takes their difference locally at each neuron. Further, the form of synaptic modification necessary to implement this algorithm is consistent with (though not directly established by) known properties of biological synaptic modification mechanisms. Finally, there are a wealth of potential sources for the necessary teaching signals in the form of actual environmental outcomes that can be compared with internal expectations to provide error signals (McClelland, 1994; O'Reilly, 1996).

Are Connectionist Models too Powerful?: Data Fitting versus Emergent Properties

To hear YB tell it, our model is bad because it cannot account for many key findings about covert recognition and normal face processing (familiarity effects, forced-choice cued recognition, provoked overt recognition, sequential associative priming, repetition priming, asymmetric interference, time to make decisions, patterns of error, and distinctiveness) and it is also bad because it can account for *anything.* At least one of these criticisms must be wrong! In our view, they both are. In the first two sections of this article we presented results that counter the first criticism. Tables 1 and 2 summarize these results. But perhaps these accomplishments do not matter, because of the second criticism, that connectionist models can simulate anything and are therefore unfalsifiable.

The claim that connectionist networks can simulate anything is true only in a very narrow, and irrelevant, sense. It is true that, given enough hidden units, a network can be trained to learn any well-defined function, that is, any set of one-to-one or many-to-one input-output mappings. In other words, networks can be explicitly trained to implement any arbitrary look-up table. But the model behaviors of interest are invariably emergent from features of its implementation, and not the result of explicit training to fit the data.

In what sense is the FOV model a look-up table? Its pre-damage ability to associate faces, semantic representations and names could be described in this way, and indeed this behavior was explicitly trained. But this is the only behavior that was explicitly trained, and it is not one of the behaviors on which any theoretical conclusions were based. In contrast, the various behavioral measures of the effects of damage on overt and covert recognition were not trained; None of the findings listed in Tables 1 and 2 were accomplished by anything analogous to table look-up in our model. Rather, these aspects of model behavior are emergent from the computational implementation of the face-semantics-name associations.

To make clear the difference between the simulations we report with FOV and the kind of simulation that is successful because of the sheer power of networks to learn any well-defined function, consider how a network would have to be trained to learn the findings listed in Tables 1 and 2. First of all, to exploit this sense of "power," one would have to somehow make variables like the degree of damage and the familiarity measure explicit components of the patterns to be learned. Then, to simulate the effects of damage on overt recognition, for example, one would train the network to map face patterns when accompanied by an explicit representation of "no damage" onto their correct semantic and name patterns, and onto an explicit representation of "familiar." Further, these face patterns when accompanied by explicit representations of different levels of damage would have to be trained to map onto partially incorrect semantic and name patterns, and onto representations of "unfamiliar." To simulate covert recognition in the matching task, for example, one would train the network to map the same face and damage input patterns to explicit representations of speed of matching, such as "1200 msec average matching time," with shorter matching speeds trained for patterns in the previously known set and longer matching speeds trained for all other patterns.

Whereas one cannot falsify the trivial approach just described, because the networks are indeed powerful enough to fit any data set given enough units and learning trials, accounts based on emergent properties are fully falsifiable. We have already referred to the tendency of networks to generalize, confuse, and prime on the basis of similarity. These are properties that emerge inexorably from interactive distributed representa-

tions. Therefore, if the human data were otherwise, the broad class of models would be disconfirmed. The same holds true for specific models such as FOV. For example, faster settling of familiar faces was not built into the model, but emerged independent of the modelers' control. If humans had shown faster perception of unfamiliar faces (a perfectly conceivable state of affairs) the model would be disconfirmed.

In sum, the allegation that connectionist models are powerful enough to simulate anything, in the context of the present debate, is based on a confusion between two kinds of power: the power to learn any well-defined function that explicitly taught, and a scientifically more interesting power to provide parsimonious explanations of complex psychological phenomena in terms of emergent properties of the computational architecture.

The explanatory power of connectionist models such as FOV lies in their emergent properties, the range of behaviors that they exhibit as a result of a relatively small set of computational principles. That these principles have some independent support from neurobiology adds to the likelihood that the models are correct. We agree with YB that the best approach to the issue of the computational architecture for cognition is not entirely general, but focuses on specific models with an eye towards the general issues. In the case of covert recognition of faces, our distributed connectionist FOV model explains all of the data that the localist IAC model explains, and more, and does so with a parsimonious and independently motivated set of principles.

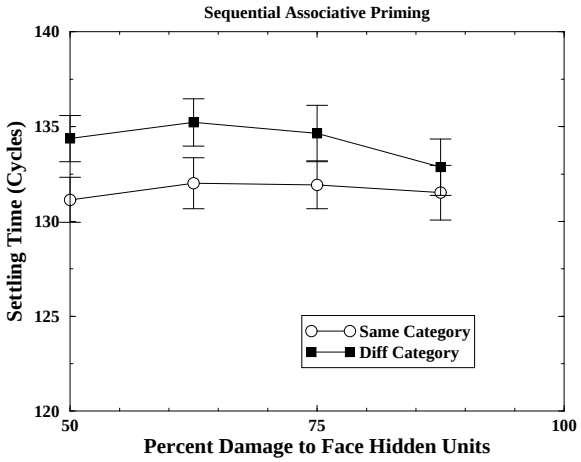

Figure 1: Sequential associative priming results, showing faster settling for names preceded by faces of people from the same category.

Acknowledgements

We would like to thank Mike Mozer for useful comments and discussion, particularly in clarifying the two senses of localist and distributed representations.

Appendix 1: Sequential Associative Priming

In this simulation the network was presented with a face input pattern and the network was allowed to settle to equilibrium. The activation states of the network were then decayed (by a factor of 90%) towards 0 (simulating fatigue),following which a name input pattern was presented, and the network again allowed to settle. The number of cycles to settle for this name pattern was the dependent measure.

There were two conditions: *same* semantic category and *different* semantic category primes. For the *same* condition, each name input was paired with a face from the same semantic category (but never with the same individual's face), and for the *different* condition, the face and name inputs were chosen from different semantic categories.

New patterns were created for this simulation because, as explained in the article, the within-category overlap of the original patterns was small (one unit) and thus allowed only very weak manipulations of semantic relatedness. The occupation subpattern of the new patterns was larger. To retain the small network size of the original model, a somewhat complicated procedure was necessary to ensure that the resulting distributed semantic representations properly captured the within-category similarity without introducing confounds that would affect the results. For example, the overlap of the face and name representations within and between categories had to be controlled.

All patterns were made with the same basic procedure: four prototype patterns were created (one for actor, one for politician, and two "other" categories). Then 10 instances of each prototype were created, by flipping some number of bits (i.e., changing a +1 to a -1 or vice versa) from the prototype, while maintaining minimum and maximum distance limits within a given category, and a minimum distance between items in different categories. For face and name inputs, the prototypes each had 5 bits active (out of 16) and a minimum hamming distance of 6, while the semantic patterns had 6 bits active (out of 18) and a minimum distance of 8. The face/name instances were produced by flipping 3 bits on and 3 bits off, with distance limits between 4 and 10 within and 4 between, which effectively eliminated any of the similarity produced

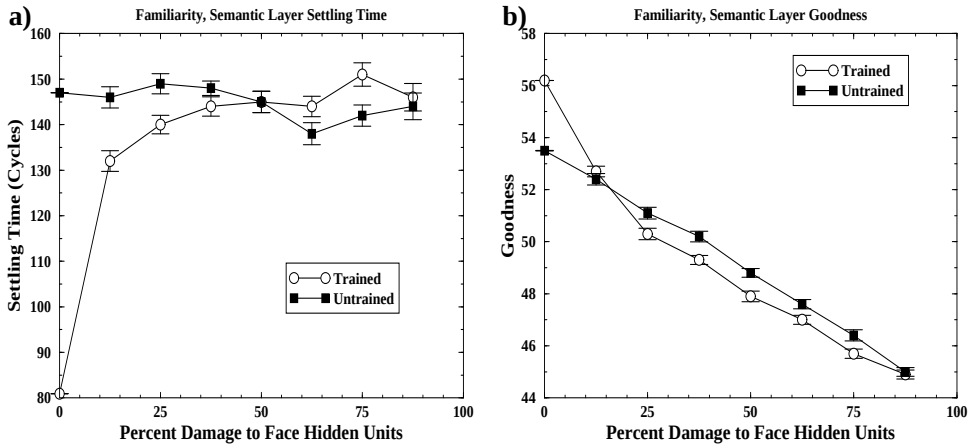

Figure 2: Familiarity results, with two different measures. **a)** shows settling time over the semantics layer (faster settling means greater familiarity), and **b)** shows goodness (negative energy) over the semantics layer (larger goodness means greater familiarity).

by the prototype. In contrast, the semantic instances flipped one bit on and one off, with distance limits of 2 and 6 within and 6 between, retaining similarity within a category. However, some amount of distance among semantic patterns is necessary for the network to learn the unique name and face mapping associated with this semantic pattern. In addition, a further check was done to ensure that the average distance of face and name patterns within a category and that between categories was essentially equal, so that these differences wouldn't create an artifactual priming effect, or obscure a true priming effect.

A network was then trained with these new patterns, and the full battery of tests as reported for the original FOV model were run (in addition to all the other tests reported here). In all cases, this network exhibited comparable performance to the original model.

The results for the sequential associative priming case are shown in Figure 1 for 100 random lesions at each of the critical levels of damage above 50% (where overt measures are at chance level performance). Note that there is a significantly faster settling time for the same condition than the different condition all the way up to 75% damage. This simulation demonstrates that sequential associative priming does indeed occur in FOV. With larger patterns allowing both more overlap and more differences, this result would be even stronger.

Appendix 2: Familiarity

In this simulation, we simply presented 10 trained face patterns ("familiar") and 10 untrained face patterns ("unfamiliar") to the network, and recorded the number of cycles necessary for the semantic units to reach our standard equilibrium criterion (maximum change in activation $< .01$). Twenty-five different random lesions were made each in increments of 12.5% of the 16 face hidden units. The results are shown in Figure 2a. Note that the difference in settling time between familiar and unfamiliar faces disappears at the 50% damage level, a level at which covert measures continue to show differences.

To compare this result with another commonly-used familiarity measure, we also recorded the goodness (negative energy) of the semantic units after the network had settled. Goodness is computed as:

$$G = \sum_i \sum_j a_i a_j w_{ij} \tag{1}$$

which gives a measure of the extent to which the activation state satisfies the constraints imposed by the

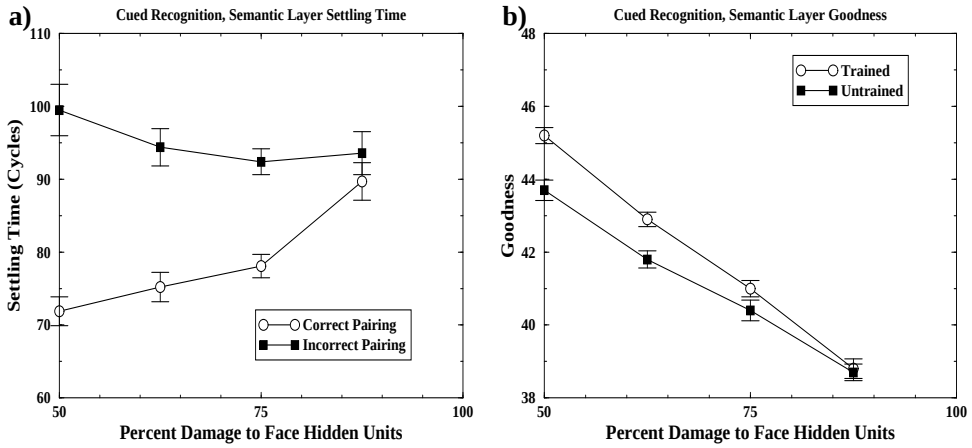

Figure 3: Forced choice cued recognition results, with two different measures of familiarity. **a)** shows settling time over the semantics layer (faster settling means greater familiarity), and **b)** shows goodness (negative energy) over the semantics layer (larger goodness means greater familiarity).

input and the weights (Hopfield, 1984; Smolensky, 1986). Greater familiarity would be associated with larger goodness values. Figure 2b shows the goodness results, which are substantially different from the settling time results in that the advantage for trained faces disappears at only 25% damage. Thus, familiarity results depend significantly on which measure is being used, leading us to be suspicious of these measures.

Appendix 3: Forced Choice Cued Recognition

Both name and face input patterns were presented to the network, and, as before, both the settling time over the semantics layer units and the goodness were used as a measures of familiarity. In one condition, the name and face were from the same person (the *correct pairing* condition), and in the other the name and face were from different people (the *incorrect pairing* condition). Again 25 random lesions were made at each level of lesion. The results are shown in Figure 3. Thus, even at 75% damage, both familiarity measures indicate that the network has a reliable preference for the correct pairing over the incorrect one. This is true even when the face input presented alone is incapable of producing either correct forced-choice naming or familiarity response.

Appendix 4: Provoked Recognition

We simulated subjects' successive fixations of multiple faces by sequentially presenting the face patterns of a category with a decay of .95. Familiarity was simulated by semantic settling time and naming accuracy by the the 10 AFC procedure, and these measures were compared following grouped presentation (as just described) and individual presentation (initializing the activations completely between presentations). Given that the accumulation of activation between same-category presentations depends on overlap in the distributed semantic representations, we used the prototype-based representations developed for the sequential associative priming simulation (described in Appendix 1). Figure 4 shows that for the the critical damage levels between 50 and 75%, familiarity and overt naming increase with the grouped presentation. The effect is small, and 500 random samples were run for each point to obtain significant results for the 10AFC measure. As discussed in Appendix 1 in connection with sequential priming, increasing the size of the network would allow for larger effects.

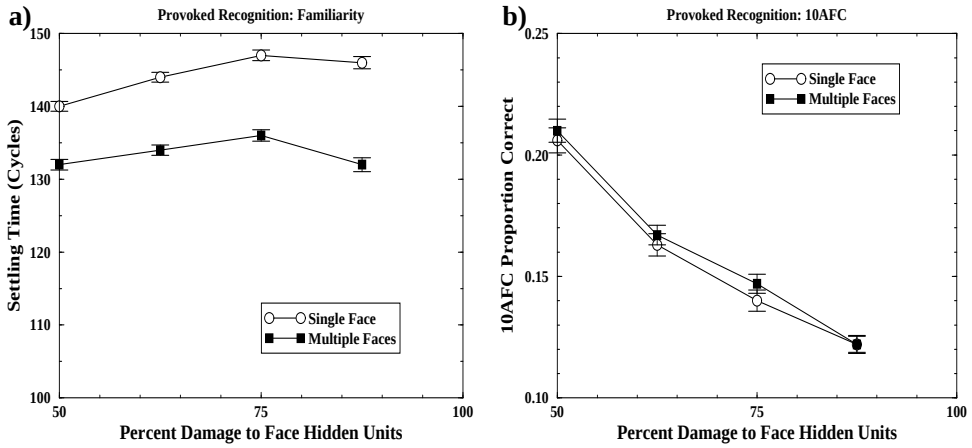

Figure 4: Provoked recognition performance of the network, comparing faces presented individually (single face) with multiple sequential faces from the same category (multiple faces). **a)** shows familiarity as indexed by setting time, and **b)** shows 10 alternative forced choice naming, 10AFC. Both measures show an advantage for multiple faces.

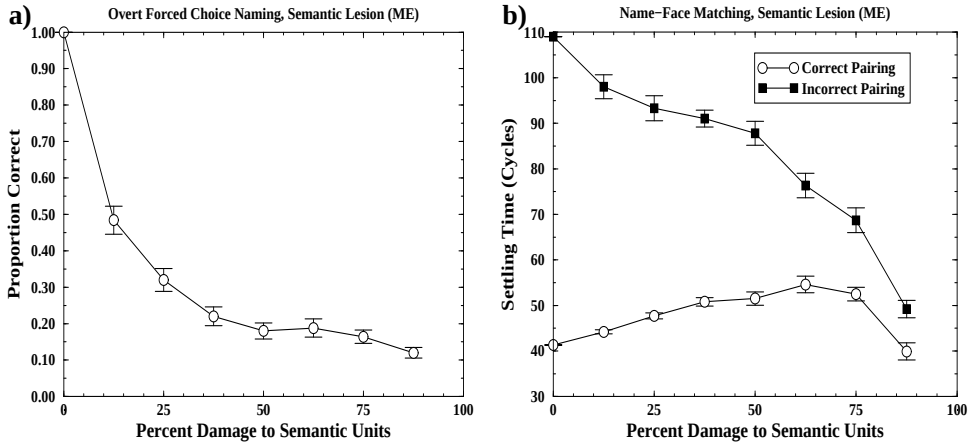

Figure 5: Performance with semantic lesions as a model of patient ME. **a)** shows the same pattern of rapidly impaired overt forced-choice face naming performance as observed in the original model. **b)** shows the name-face matching performance (i.e., cued recognition) for these same levels of damage, which is preserved up to high levels of damage.

Appendix 5: Semantic Lesions (Patient ME)

In order to simulate patient ME, who cannot retrieve semantic information about familiar people, we damaged the semantic layer of our network, running 25 different lesions at each lesion level. As a demonstration of the dramatically impaired overt performance that results from this damage, we show the same pattern of highly impaired overt forced-choice face naming performance in Figure 5a. However, even at the higher levels of damage where this overt performance is extremely impaired, the name-face matching performance (instantiated as in the cued-recognition simulation described above) remains very good, as shown in Figure 5b.

References

Andersen, R. A., & Zipser, D. (1988). The role of the posterior parietal cortex in coordinate transformations for visual-motor integration. *Canadian Journal of Physiological Pharmacology, 66*, 488–501.

Becker, S., Moscovitch, M., , Behrmann, M., & Joordens, S. (1997). Long-term semantic priming: a computational account and empirical evidence. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 1059–1082.

Borowsky, R., & Masson, M. E. J. (1996). Semantic ambiguity effects in word identification. *Journal of Experimental Psychology: Learning Memory and Cognition, 22*(1), 63–85.

Burgess, N. (1995). A solvable connectionist model of immediate recall of ordered lists. In D. S. Touretzky, G. Tesauro, & T. K. Leen (Eds.), *Advances in neural information processing systems* (pp. 51–58). Cambridge, MA: MIT Press.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing model of the stroop effect. *Psychological Review, 97*(3), 332–361.

Dayan, P. (1998). A hierarchical model of binocular rivalry. *Neural Computation, 10*, 1119–1136.

de Haan, E. H., Young, A. W., & Newcombe, F. (1987). Face recognition without awareness. *Cognitive Neuropsychology, 4*, 385–415.

Desimone, R., & Ungerleider, L. G. (1989). Neural mechanisms of visual processing in monkeys. In F. Boller, & J. Grafman (Eds.), *Handbook of neurophysiology, vol. 2* (Chap. 14, pp. 267–299). Amsterdam: Elsevier.

Farah, M. J. (1990). *Visual agnosia.* Cambridge, MA: MIT Press.

Farah, M. J., O'Reilly, R. C., & Vecera, S. P. (1993). Dissociated overt and covert recognition as an emergent property of a lesioned neural network. *Psychological Review, 100*, 571–588.

Georgopoulos, A. P. (1990). Neurophysiology and reaching. In M. Jeannerod (Ed.), *Attention and performance, vol. 13* (pp. 227–263). Hillsdale, N.J.: Erlbaum.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In D. E. Rumelhart, J. L. McClelland, & PDP Research Group (Eds.), *Parallel distributed processing. volume 1: Foundations* (Chap. 3, pp. 77–109). Cambridge, MA: MIT Press.

Hinton, G. E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review, 98*(1), 74–95.

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences, 81*, 3088–3092.

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

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, J. T., & Xu, F. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development, 57*(4), 1–165.

Mathis, D. A., & Mozer, M. C. (1996). Conscious and unconscious perception: A computational theory. In G. Cottrell (Ed.), *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society* (pp. 324–328). Hillsdale, NJ: Erlbaum.

Mayall, K., & Humphreys, G. (1996). A connectionist model of alexia: Covert recognition and case mixing effects. *British Journal of Psychology, 87*, 355–402.

McClelland, J. L. (1994). The interaction of nature and nurture in development: A parallel distributed processing perspective. In P. Bertelson, P. Eelen, & G. D'Ydewalle (Eds.), *Current advances in psychological science: Ongoing research* (pp. 57–88). Hillsdale, NJ: Erlbaum.

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 connectionst models of learning and memory. *Psychological Review, 102*, 419–457.

Metcalfe, J., Cottrell, G. W., & Mencl, W. E. (1992). Cognitive binding: A computational-modeling analysis of a distinction between implicit and explicit memory. *Journal of Cognitive Neuroscience, 4*, 289–298.

Munakata, Y., McClelland, J. L., Johnson, M. J., & Siegler, R. S. (1997). Rethinking infant knowledge: Toward an adaptive process account of successes and failures in object permanence tasks. *Psychological Review, 104*, 686–713.

O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation, 8*(5), 895–938.

Plaut, D. C. (1997). Structure and function in the lexical system: Insights from distributed models of word reading and lexical decision. *Language and cognitive processes, 12*, 767–808.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. E. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103*, 56–115.

Rolls, E. T. (1989). Functions of neuronal networks in the hippocampus and neocortex in memory. In J. H. Byrne, & W. O. Berry (Eds.), *Neural models of plasticity: Experimental and theoretical approaches* (pp. 240–265). San Diego, CA: Academic Press.

Rueckl, J. G. (1995). Ambiguity and connectionist networks: Still settling into a solution–Comment on Joordens and Besner (1994). *Journal of Experimental Psychology: Learning Memory and Cognition, 21*(2), 501–508.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96*, 523–568.

Seidenberg, M. S., Waters, G. S., Sanders, M., & Langer, P. (1984). Pre- and postlexical loci of contextual effects on word recognition. *Memory and Cognition, 12*, 315–328.

Shuttleworth, E. C., Syring, V., & Allen, N. (1982). Further observations on the nature of prosopagnosia. *Brain and Cognition, 1*, 307–322.

Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In D. E. Rumelhart, J. L. McClelland, & PDP Research Group (Eds.), *Parallel distributed processing. volume 1: Foundations* (Chap. 5, pp. 282–317). Cambridge, MA: MIT Press.

Sparks, D. L., & Mays, L. E. (1990). Signal transformations required for the generation of saccadic eye movements. *Annual Review of Neuroscience, 13*, 309–336.

Young, A. W., & Burton, A. M. (in press). Simulating face recognition: Implications for modelling cognition. *Cognitive Neuropsychology*.

Young, A. W., Newcombe, F., de Haan, E. H., Small, M., & Hay, D. C. (1993). Face perception after brain injury. *Brain, 116*, 941–959.

Young, M. P., & Yamane, S. (1992). Sparse population coding of faces in the inferotemporal cortex. *Science, 256*, 309–316.