A Connectionist Account of Asymmetric Category Learning in Early Infancy

To appear in *Developmental Psychology*, (2000).

Denis Mareschal
Centre for Brain
& Cognitive Development
Birkbeck College
dmareschal@bbk.ac.uk

Robert M. French
Université de Liege
rfrench@ulg.ac.be

Paul Quinn
Brown University

Running Head: Asymmetric category learning

This work was supported in part by a collaborative research grant awarded to the first two authors by the British Council and the Belgian CGRI, by Belgian FNRS Grant No. D.4516.93 and PAI Grant No. P4/19 awarded to the second author, and by Grant BCS-9816641 from the National Science Foundation to the third author. We are grateful to Les Cohen, James Dannemiller, Carolyn Rovee-Collier, Tom Shultz, and 3 anonymous reviewers for helpful comments on an earlier draft of this article.

Address all correspondence to Denis Mareschal, Centre for Brain and Cognitive Development, Department of Psychology, Birkbeck College, University of London, Malet Street, London, WC1E 7HX, UK. Email: d.mareschal@bbk.ac.uk

Abstract

Young infants show unexplained asymmetries in the exclusivity of categories formed on the basis of visually presented stimuli. We describe a connectionist model that shows similar exclusivity asymmetries when categorizing the same stimuli presented to the infants. The asymmetries can be explained in terms of an associative learning mechanism, distributed internal representations, and the statistics of the feature distributions in the stimuli. We use the model to explore the robustness of this asymmetry. The model predicts that the asymmetry will persist when a category is acquired in the presence of mixed category exemplars. A study with 3- to 4-month-olds show that asymmetric exclusivity continues to persist in the presence of a mixed familiarization, thereby corroborating the model's predictions. We suggest that by interpreting asymmetric exclusivity effects as manifestations of interference in an associative memory system, the model can also be extended to account for interference effects in early infant visual memory.

A Connectionist Account of Asymmetric Category Learning in Early Infancy

Young infants can form perceptual category representations when presented with a set of perceptually similar stimuli from the same class. These representations allow infants to organize their perceptual experiences into groupings that in many instances come to have conceptual significance for children and adults. For example, by 3 or 4 months of age infants have been shown to categorize a range of real world images of Cats, Dogs, Horses, Chairs, and Couches (Madole & Oakes, 1999; Quinn, 1998; Quinn & Eimas, 1996a). However, the perceptual category representations do not always have the same characteristics as might be expected from the corresponding adult category representations. In particular, the extension and exclusivity of the perceptual category representations of infants (i.e., the range of exemplars accepted or rejected as members of the category) may differ from those of adult category representations.

Quinn, Eimas, and Rosenkrantz (1993) used a familiarization/novelty-preference technique to determine if the perceptual category representations of familiar animals (e.g., cats and dogs) acquired by young infants would exclude perceptually similar exemplars from contrasting basic-level categories. They found that when 3- to 4-month-olds are familiarized with six pairs of cat photographs presented sequentially (12 photographs), the infants will subsequently prefer to look at a novel dog photograph rather than a novel cat photograph. Because infants have an inherent preference to look at unfamiliar stimuli (Fagan, 1970; Fantz, 1964; Slater, 1995), this result was interpreted as showing that the infants had developed a category representation of Cat that included novel cats (hence less looking at the cat photograph) but excluded novel dogs (hence more looking at the dog photograph). However, if the infants are initially familiarized with six pairs of dog photographs sequentially (12 photographs), they will show no subsequent preference for looking at either a novel dog or a novel cat. Furthermore, control conditions revealed that: (1) the infants would prefer to look at a novel test bird after initial familiarization with either dogs or cats, (2) there is no a priori preference for dogs over cats, and (3) the infants are able to discriminate within the Cat and Dog categories. Taken together, these findings led Quinn et al. to suggest that the 3- to 4-month-olds had formed a perceptual category representation of Dog that included novel dogs but also included novel cats.

There appears to be an asymmetry in the exclusivity of the two perceptual category representations formed during familiarization. The Cat representation excludes novel dogs, whereas the Dog representation does not exclude novel cats. The reason for this asymmetry remains unclear, although Quinn et al. (1993) presented some evidence that it might be related to greater variability among dogs. We believe that a full explanation of this asymmetry requires a mechanistic account of how categories are formed by infants during a test session. Some researchers have tried to identify what information within a set of stimuli might be used by infants to delimit a perceptual category representation (Quinn & Eimas, 1996b; Younger, 1985, Younger & Cohen, 1986). Although this approach can be very revealing about the basis for categorization, it leaves unanswered the question of how categories are causally derived from the set of exemplars experienced by the infant.

Given that the role of developmental psychology is to establish the causal mechanisms by which behavior emerges, we propose in this article to explore HOW and WHY young

infants categorize complex visual images in the way they do. To achieve this goal we will present a combination of connectionist (computational) modeling and experimental studies of infant behaviors designed to test the legitimacy of the modeling.

Computational modeling provides a tool for exploring the mechanisms that underlying behavior. Connectionist models are computer models loosely based on the principles of neural information processing (Elman, Bates, Johnson, Karmiloff-Smith, Parisi, & Plunkett, 1996; Mareschal, in press; McLeod, Plunkett, & Rolls, 1998; Rumelhart & McClellend, 1986). They are not intended to be neural models. Instead, they attempt to strike the balance between importing some of the basic concepts from the neurosciences while formulating questions about behavior in terms of high-level cognitive concepts.

From a developmental perspective, connectionist networks are ideal for modeling because they develop their own internal representations as a result of interacting with a structured environment (Plunkett & Sinha, 1991; Mareschal & Shultz, 1996). Although connectionist modeling has its roots in associationist learning paradigms, it has inherited the Hebbian rather than the Hullian tradition. That is, what goes on inside the network (i.e., the internal representation of information) is as important in determining the overall behavior of the network as are the correlations between the inputs (stimuli) and the outputs (responses).

Building a model of infant categorization

Many infant visual categorization tasks rely on preferential looking techniques based on the finding that infants direct attention more to unfamiliar or unexpected stimuli (Fagan, 1970; Fantz, 1964; Slater, 1995). The standard interpretation of this behavior is that the infants are comparing an input stimulus to an internal representation of the same stimulus (e.g., Charlesworth, 1969; Cohen, 1973; Sokolov, 1963). As long as there is a discrepancy between the information stored in the internal representation and the visual input, the infant continues to attend to the stimulus. While attending to the stimulus the infant updates its internal representation. When the information in the internal representation is no longer discrepant with respect to the visual input, attention is switched elsewhere. When a familiar object is presented there is little or no attending because the infant already has a reliable internal representation of that object. In contrast, when an unfamiliar or unexpected object is presented, there is much attending because an internal representation has to be constructed or adjusted. The degree to which the novel object differs from the information stored in the existing internal representations determines the amount of adjusting that has to be done, and hence the duration of attention.

We used a connectionist autoencoder to model the relation between attention and representation construction (cf. Mareschal & French, 1997; Mareschal & French, 2000; Schafer & Mareschal, in press). An autoencoder is a feed-forward connectionist network with a single layer of hidden units (Ackley, Hinton, & Sejnowski, 1985; Rumelhart & McClelland, 1986). The network learns to reproduce on the output units the pattern of activation across the input units. The number of hidden units must be smaller than the number of input or output units. This architectural constraint produces a bottleneck in the flow of information through the network. Learning in an autoencoder consists of developing a more compact internal representation of the input (at the hidden unit level)

that is sufficiently reliable to reproduce all the information in the original input. Hence the incentive to develop category-based representations. Information is first compressed into an internal representation and then expanded to reproduce the original input. The successive cycles of training in the autoencoder are an iterative process by which a reliable internal representation of the input is developed. The reliability of the representation is tested by expanding it, and comparing the resulting predictions to the actual stimulus being encoded. Similar networks have been used to produce compressed representations of video images (Cottrell, Munro, & Zipser, 1988).

We suggest that during the period of captured attention infants are actively involved in an iterative process of encoding visual input into an internal representation and then assessing that representation against continuing perceptual input. This is accomplished by using the internal representation to predict what the properties of the stimulus are. As long as the representation fails to predict the stimulus properties, the infant continues to fixate the stimulus and to update the internal representation. Similar interpretations have been suggested elsewhere (Mareschal & French, 2000; Mareschal, Plunkett, & Harris, 1999; Munakata, McClelland, Johnson, & Siegler, 1997; see also Di Lollo, Enns, & Rensink, in press, for a comparable account of adult visual object recognition).

This modeling approach has several implications. It suggests that infant looking times are positively correlated with network error[1]. The greater the error, the longer the looking time. Stimuli presented for a very short time will be encoded less well and produce more error than those presented for a longer period. However, prolonged exposure after error (attention) has fallen off will not improve memory of the stimulus. The degree to which error (looking time) increases on presentation of a novel object depends on the similarity between the novel object and the familiar object. Presenting a series of similar objects from the same perceptual category leads to a progressive error drop on future similar objects. A prototype of the set of objects leads to lower error than individual objects. All of this is true of both autoencoders (where output error is the measurable quantity) and infants (where looking time is the measurable quantity).

Simulation 1: The development of Cat and Dog categories

The modeling results in this section and all subsequent simulation sections are based on the performance of a standard 10-8-10 feed-forward backpropagation network[2]. Autoencoders are reasonably robust to variations in the number of hidden units and the value of specific parameters. One requirement for efficient autoencoding is that there is a sufficient number of hidden units to capture the principal components of variation in the data. However, too many hidden units may reduce the network's ability to generalize to novel exemplars. These, and other aspects of autoencoding, are discussed in Hertz, Krogh, and Palmer (1991).

To model the original exclusivity asymmetry effect, data for training the networks were obtained from measurements of the original Cat and Dog pictures used by Quinn et al. (1993). These data are tabulated in the Appendix. There were 18 dogs and 18 cats classified according to the following ten traits: head length, head width, eye separation, ear separation, ear length, nose length, nose width, leg length, vertical extent, and horizontal extent. Although it is difficult to say for certain which features the infants

are using during categorization, it is well known that infants can segregate items into categories on the basis of attributes with different values (Younger, 1985; see Quinn & Johnson, 1997, for a detailed justification of similar input features). The feature values were normalized[3] to be within 0 and 1. The input data are discussed in more detail below.

Twelve items from one category were presented sequentially to the network in groups of two (i.e., weights were updated in batches of two) to capture the fact that pairs of pictures were presented to the infants during the familiarization trials. Networks were trained for 250 epochs (weight updates) on one pair of patterns before being presented with the next pair. This was done to reflect the fact that in the Quinn and Eimas studies infants were shown pairs of pictures for a fixed duration of time. The total amount of training was, therefore, 6x250=1500 weight updates. The results are averaged over 50 network replications, each with random initial weights. The remaining six items from each category were used to test whether the networks had formed category representations.

Like infants, these networks form both Cat and Dog categories. Figure 2 shows the initial error score (i.e., the sum-squared-error between the actual output produced and ideal target value, measured across all output units), the error score after twelve presentations of either cats or dogs, and the average error score (after training) for the 6 remaining exemplars in either the Cat or Dog category. After learning, error is lower, suggesting that the network has developed reliable internal representations of cats or dogs. The generalization error rises slightly, showing that the networks recognize these exemplars as novel. Infants are also able to distinguish individual exemplars within the category (Quinn et al., 1993). However, the generalization error remains well below the initial error suggesting that the new exemplars are assimilated within the category representation formed by the networks across the hidden units.

========== Insert figure 2 about here ==========

The Asymmetric Exclusivity of the Cat and Dog Categories

Eimas and Quinn found that there was an asymmetry in the exclusivity of the Cat and Dog categories developed by infants. Figure 3 shows what happens when networks trained on cats are presented with a novel cat and a novel dog, and when networks trained on dogs are tested with a novel dog and a novel cat. In these network models (as with the infants) the acquisition of categorical representations is inferred from their differential sum-squared-error response (preferential looking responses) when presented with a novel exemplar of the familiar category or a novel exemplar of the a novel category. The comparison of these error scores relative to each other reflects the relative looking time towards each stimulus observed with the infants.

When the networks are initially trained on cats, the presentation of a novel dog results in a large error score relative to that produced by the presentation of a novel cat exemplar, corresponding to the results observed with infants in terms of a longer looking time. Dogs are not included within the category representation of cats. In contrast, when the networks are initially trained on dogs, the presentation of a novel cat results in only a small increase in error relative to that produced by the presentation of a novel dog, suggesting that the cats have been included in the dog category. Hence, the networks also show an asymmetry in the exclusivity of the category representations developed.

========== Insert Figure 3 about here ==========

It could be argued that the asymmetry arises from the unequal learning of the Cat and Dog categories by the networks. It may take longer to learn the Dog category than the Cat category; hence, training for a fixed number of epochs would lead to a less established representation of Dog by the networks. To address this possibility, 50 new networks were trained as above but with one exception. These networks were trained to a fixed error criterion rather than a fixed epoch criterion. Networks were trained on each pair of familiarization exemplars until all output units were within 0.2 of their target values[4]. This ensures that the networks have learned to autoencode every input to the same minimum criterion. Under these conditions, networks familiarized with cats showed an average error of 0.26 (SD=0.05) and 0.45 (SD=0.03) when presented with a novel cat and a novel dog respectively, whereas networks familiarized with dogs showed an average error of 0.35 (SD=0.15) and 0.41 (SD=0.12) when presented with a novel dog and a novel cat respectively. Thus, the exclusivity asymmetry persists even when networks are familiarized to a fixed error criterion.

Distribution of Features in the Stimuli

The associative learning mechanisms embodied in connectionist networks (and described by the mathematics of these networks), when coupled with the non-linear response of hidden units provides a mechanistic account of HOW information is processed in such a system. However, a full explanation of the asymmetric exclusivity requires an account of WHY the system learns a representation for cats that excludes dogs, but learns a representation for dogs that includes cats. Connectionist networks extract the correlations between features present in their learning environment. The distibutional characteristics of the internal representations (developed across the hidden units) reflect the distibutional characteristics of the corresponding categories in the environment. This suggests that an explanation for why the networks exhibit an exclusivity asymmetry may be found by examining the input data. Figure 4 shows the probability distributions of the 10 traits for both cats and dogs, when fit to gaussian distributions with means and standard deviations derived from the normalized feature values[5]. Some of the traits are very similar in terms of their means and distribution for both cats and dogs (e.g., head length and head width). Others, especially nose length and nose width, are very different and will provide the crucial explanation of the unexpected looking asymmetries reported by Quinn et al. (1993).

========== Insert Figure 4 about here ==========

Consider the single trait nose width. The (normalized) mean nose width for the dog population is 0.53 with a standard deviation of 0.20, whereas the mean for the cat population is 0.24 with a much smaller standard deviation of 0.07. Consequently, the nose width of virtually all cats in the population will fall within two standard deviations of the nose-width mean for dogs. On the other hand, the nose width of the majority of dogs does not fall within two standard deviations of the nose-width mean for cats. The

result, in short, is that at least for this trait, all cats could be exemplars of dogs, whereas most dogs could not be cats.

When we examine all of the members of the two populations, we see that the values of all 10 traits for 9 (i.e., 50%) of the members of the Cat category fall within a two standard deviation cut-off for those traits for the Dog category. Fully half of the cats in the population could be reasonably classified as dogs. In contrast, the smaller means and lower variances of a number of traits (especially, nose length and nose width) for cats compared to dogs means that only 2 of the 18 dogs (i.e., 11%) could conceivably be classified as members of the Cat category.

A corollary of this finding is that networks will, on average, generalize their autoencoding responses better in the dog-to-cat direction than in the cat-to-dog direction. When presented with a novel cat, a network trained on dogs will recognize this item as a member of the category it has learned (i.e., DOG), and for which it has also learned to output an appropriate feature description. In contrast, when presented with a novel dog, a network trained on cats will fail to recognize this item as a member of the category it has learned (i.e., CAT), and will be unable to output an appropriate feature description.

The exclusivity asymmetry of the categories formed on the basis of these exemplars reflects the distribution of features characteristic of the cat and dog exemplars presented to the networks (and infants). The key feature of the data is that the distribution of values for the Cat features are (in general) subsumed within the distribution of Dog features. There are components to these distributional characteristics: (1) the Dog distributions are sometimes broader than the Cat distributions, and (2) the Cat values are often included within the range of Dog values. The greater range of Dog values is a necessary but not sufficient condition ofr the asymmetric exclusivity to appear. If, on the one hand, the Dog and Cat distributions had the same range, then there would be no exclusion in either direction. If, on the other hand, the Cat and Dog feature values had different ranges but no overlap, then there would be an exclusivity in both directions. To see this, suppose that the cats had feature values ranging on the interval [0, 0.25] while the dogs had feature values ranging on the interval [0.5, 1.0]. Networks initially trained with cats would have no experience of processing clusters of values in the range [0.5, 1.0]. Thus, when presented with a novel dog input, they would mostly likely produce a default value of the mean of their experiences (0.125) or the maximum of their range (0.25) along each dimension. This would result in a large error observed in response to a novel dog input. Similarly, networks initially trained on dogs would have no experience of processing clusters of values in the range [0, 0.25]. When presented with a novel cat input, they would mostly likely produce a default value of the mean of their experiences (0.75) or the minimum of their range (0.5) along each dimension. This would result in a large error observed in response to a novel dog input. With no overlap in feature distributions, behavior consistent with symmetric exclusivity would be observed.

The degree to which the inclusion relationship observed in the nose width feature holds across the set of features determines the degree to which an asymmetric exclusivity effect is observed. It is not just the greater variability of dogs along certain feature dimensions that causes the exclusivity asymmetry. The inclusion relationship with respect to cats also plays a crucial role.

However, the asymmetry inherent in the data is only translated into corresponding behavior because connectionist networks (and presumably infants) develop internal

representations that reflect the distributions of the input features. Thus, the internal representation for Cat will be subsumed within the internal representation for Dog along several dimensions. It is because the internal representations share this inclusion relationship that an asymmetry in error (looking time) is observed.

Simulation 2: Learning from mixed exemplars

Categories are rarely acquired in isolation. Infants engaged in casual observation of their nursery environment presumably do not encounter numerous objects from the same natural kind or artifactual category presented in such close temporal proximity. It would be more likely for infants to encounter multiple objects from various categories in a quick scan of their immediate surroundings. The question that arises for researchers interested in the early development of categorization is how infant categorization performance will be affected when infants are presented with instances from two or more categories in the same familiarization session. It could be reasoned that experiencing two categories during learning will enhance the formation of distinct categories because examples of one category will provide a contrasting reference for the learning the other category. Alternatively, it could be argued that presenting two categories simultaneously will create interference between them thereby making the construction of a representation for each more difficult.

The evidence that is relevant to the contrast and interference hypotheses is mixed. For example, in studies of the acquisition of dot pattern categories, the category representations formed by both adults and young infants were enhanced in experimental sessions in which multiple categories were presented (Homa and Chambliss, 1975; Quinn, 1987; see also Younger, 1985, for consistent findings obtained with schematic animal stimuli). In these studies, the facilitative effect of multiple category presentation was attributed to the fact that it provided participants with the opportunity to observe both the similarities among members within a category and differences between members of different categories. However, there are also data to indicate that when more naturalistic exemplars are used as stimuli (i.e., realistic photographs of animals), then multiple category presentation results in category representations that are either not different from or less differentiated than those formed during single category presentation (Eimas & Quinn, 1994; Oakes, Plumert, Lansink, & Merryman, 1996;Younger & Fearing, 1999).

To explore how connectionist networks would behave under conditions of multiple category presentation, autoencoder networks identical to those in simulation 1 were exposed to an interleaved set of cat and dog exemplars. Fifty previously untrained networks were exposed to 8 cats and 4 dogs (the mainly-cat condition) and 50 previously untrained networks were exposed to 8 dogs and 4 cats (the mainly-dog condition). The training procedure was identical to that in simulation 1 with one exception. As before, the networks were presented with 6 pairs of exemplars over 6 familiarization trials. Networks were trained for a fixed 250 epochs with each familiarization pair. However, on four of the 6 familiarization trials, one of the exemplars in the pair was taken from the contrasting category. For example, a network in the mainly-dog condition might experience the following series of familiarization pairs: cat1-dog1, dog2-dog3, dog4-cat2, cat3-dog5, dog6-dog7 dog8-cat4. In total, these networks experienced 8 exemplars

from the dominant category and 4 exemplars from the contrasting category. The trials on which an exemplar from the contrasting category was presented were randomly selected. Networks were then tested with the remaining unfamiliar exemplars from the Cat and Dog categories and their responses were recorded.

========== Insert Figure 5 about here ==========

Figure 5 shows the networks' response to a novel cat and a novel dog after having been familiarized in either the mainly-cat or the mainly-dog conditions. In contrast to the proposal that experiencing exemplars from both categories during learning might enhance the separation of the categories, these networks showed the same asymmetric exclusivity effect as those trained only with cats or only with dogs. Networks familiarized with 8 cats and 4 dogs show much greater error when presented with a novel dog than a novel cat suggesting that they have formed a category of Cat that excludes dogs. The networks familiarized with 8 dogs and 4 cats show little difference in error when presented with novel dogs and novel cats suggesting that they have formed a category representation that includes both cats and dogs.

It could be that this asymmetry is due to the fact that the Dog category is more difficult to learn than the Cat category. However this is not the case as networks trained to a fixed error criterion rather than a fixed epoch criterion showed the same asymmetry. The networks trained to a fixed error criterion in the mainly-cat condition had a mean error of 0.32 (SD=0.02) and 0.39 (SD=0.02) when presented with a novel cat and a novel dog respectively. The networks trained to a fixed error criterion in the mainly-dog condition had a mean error of 0.39 (SD=0.02) and 0.38 (SD=0.02) when presented with a novel cat and a novel dog respectively.

The persistence of the asymmetry constitutes an explicit prediction of infant behaviors that derives from the associationist mechanisms of connectionist networks coupled with the characteristics of the stimuli used to familiarize infants, and provides a direct test of the model.

Experiment 1: Infant responses to mixed exemplar familiarization

The model makes two specific predictions. The first is that 3- to 4-month-olds familiarized according to the procedures of the mainly-cat condition will show a significant preference for novel dogs over novel cats in a subsequent preferential looking test. The second prediction is that 3- to 4-month-olds familiarized according to the procedures of the mainly-dog condition will show no preference for novel dogs over novel cats in a subsequent preferential looking test. These predictions were tested using the same cat and dog pictures encoded for simulations 1 and 2.

Method

Participants. Forty-eight 3- to 4-month-olds (27 boys, 21 girls) were participants (mean age = 103 days; SD = 15 days). Nine additional infants were not included in the analyses because of fussiness (n=7), a position bias (> 95% looking to one side of the display (n=1), or a failure to look at both test stimuli (n=1).

Stimuli. The stimuli were thirty-six color photographs of cats and dogs (18 exemplars of each category) previously used in Quinn et al. (1993) and Eimas, Quinn,

and Cowan (1994). The pictures were cut from Simon and Schuster's Guide to Cats (Siegal, 1983) and Simon and Schuster's Guide to Dogs (Schuler, 1980), and chosen to represent a variety of shapes, colors, and stances of both categories of animals. Each picture contained a single animal that had been cut away from its background and mounted onto a white 17.7 by 17.7-cm posterboard for presentation.

Apparatus. Infants were tested by means of a portable visual preference apparatus, adapted from that used by Fagan (1970). The apparatus is an enclosed viewing box with a gray display stage (85 cm wide and 29 cm high) that contains two compartments to hold the two posterboard stimuli. The stimuli were illuminated by a 60 Hz fluorescent lamp that was shielded from the infant's view. The center-to-center distance between the two compartments was 30.5 cm. A 0.625-cm peephole located midway between the stimulus compartments permitted observation and recording of the infant's visual fixations.

Procedure. The infants were tested individually. They were brought to the third author's laboratory by a parent and placed in a reclining position on the seated parent's lap. An experimenter wheeled the apparatus over the infant, keeping the infant's head centered with respect to the middle of the display stage. As soon as the infant was properly aligned and apparently at ease, a trial was begun. The experimenter loaded the stimuli from a nearby table into the stimulus compartments, elicited the infant's attention and closed the stage, thereby exposing the stimuli to the infant. The center of the display stage was approximately 30.5 cm in front of the infant while the stimuli were being viewed. During a trial, the experimenter observed the infant through the peephole and recorded fixations to the left and right stimuli using a 605 XE Accusplit stopwatch held in each hand. The criterion for fixation was observing corneal reflection of the stimulus over the infant's pupil. Interobserver reliability, as determined by comparing the looking times measured by the experimenter using the center peephole, and additional observers using peepholes to the left of the left stimulus compartment and to the right of the right stimulus compartment, was high (Pearson $r = 0.97$). This reliability was derived from observations made by independent observers on 48 novelty-preference test trials from 24 infants. This value is comparable to estimates of interobserver reliability obtained in other laboratories measuring visual fixation duration with the corneal reflection procedure (Haaf, Brewster, deSaint Victor, & Smith, 1989; O'Neill, Jacobson, & Jacobson, 1994). Between trials, the experimenter opened the stage, recorded the looking time data on a data sheet, changed the stimuli (or their position), recentered the infant's gaze, and closed the stage, thereby beginning the next trial. Two experimenters were used to record fixations, one during familiarization and another during test trials. Both were trained research assistants who were naive to the hypotheses of the studies. The experimenter recording during test trials was also naive to the stimulus information that the infant was shown during the familiarization trials.

The 48 infants were randomly assigned to one of two category presentation orders. Each infant in the mainly-cat group was first familiarized with 8 cats and 4 dogs, with exemplars of the contrasting category interleaved on four separate trials. The cats and dogs were randomly selected and different for each infant, presented during six 15-s trials (two different animals per trial). After this first familiarization phase infants were presented with a novel cat and a novel dog for two 10-s test trials. Left-right locations were counterbalanced across infants on the first test trial and reversed on the second test

trial. Infants in the mainly-dog group were familiarized and tested in the same way as those in the mainly-cat group except that they saw 8 dog pictures interleaved with 4 cat pictures.

Results

Familiarisation trials. Table 1 shows the mean fixation times averaged across the first three familiarization trials, and the second three familiarization trials. An ANOVA with condition (MAINLY-CAT vs. MAINLY-DOG) by trial block (1-3 vs. 4-6) revealed no significant effect of trial block on the initial familiarization trials ($F(1, 46)=0.05$, $p> .20$). Neither the effect of condition nor the interaction of condition and trial block were reliable, $F(1, 46)<2.50$, $p>.12$, in both cases. These findings suggest that any differences in the preference test outcomes cannot be attributed to category-specific differential habituation rates.

========== Insert Table 1 about here ==========

Preference test trials. Each infant's looking time to the stimulus from the novel category was divided by the total looking time to both stimuli and converted to a percentage score. Mean category preference scores are shown in Table 2.

========== Insert Table 2 about here ==========

The model predicts that infants familiarized in the mainly-cat condition will show a significant preference for a novel dog exemplar. The model also predicts that infants familiarized in the mainly-dog condition will show no preference for the novel cat or the novel dog. Inspection of Table 2 shows that these predictions have been confirmed with 3- to 4-month-old infants. The mean novel category preference for novel dogs in the mainly cat condition was significantly higher than chance, whereas the mean novel category preference for cats in the mainly dog condition was not different from chance. In addition, the novel category preferences for the mainly cat and mainly dog conditions were reliably different from each other, $t(47) = 2.08$, $p < .05$, two-tailed.

In summary, these results support the model's predictions. Interleaved learning was not found to consolidate the formation of two distinct categories. The asymmetry in novelty preference to cat and dogs, identified in the original Quinn and Eimas studies, persists even when infants are exposed to a small number of interleaved exemplars of a contrasting category during familiarization.

General Discussion

Like young infants, connectionist autoencoder networks formed categorical representations of cats and dogs when presented with the same stimuli as the infants. The category representations showed asymmetric exclusivity. The Cat category included novel cat exemplars but excluded novel dog exemplars, whereas the Dog category included novel dog exemplars as well as novel cat exemplars. The category asymmetry was suggested to be related to the distribution of features in the stimuli shown to the infants. More specifically, the asymmetry arose because the connectionist networks developed internal representations reflecting the overlap in the distribution of the features in the two sets of stimuli. Most of the cat exemplars could be classified as dogs, whereas most dogs were not plausible cats. The asymmetric category representations reflect an interaction between the statistics of the learning environment (the images shown to the infants) and the computational properties of an associative learning system with

distributed representations. Infant performance in these categorization tasks is essentially driven by a bottom-up process.

The exclusivity asymmetry was also explored by investigating whether the addition of exemplars from a contrasting category would enhance the formation of distinct Cat and Dog categories. Connectionist networks trained with 8 exemplars from one category and 4 exemplars from a contrasting category continued to show asymmetric exclusivity in the categories developed. These results constituted a set of predictions about the behavior of 3- to 4-month-olds that enabled an evaluation of the model. Infants familiarized with the same mixture of cat and dog exemplars were found to show an asymmetry in their responses to novel cat and dog exemplars that was consistent with the model.

The new evidence reported in Experiments 1 strongly supports the connectionist account of early infant categorization. The asymmetry can be explained by appealing to the associative learning mechanisms of connectionist networks and the statistical distribution of features in the stimuli used to familiarize infants. The connectionist mechanisms account for how the behavior emerges and the input data account for why the behavior emerges in the presence of these cat and dog pictures. It could be argued that the analysis of the data alone provides the explanation of infant behaviors and that the network account contributes little to this explanation. However, this is not the case. One can see this by trying to make predictions about how an unknown "black-box" might respond to this same data. Without any information about the mechanisms that translate input to observed behaviors within this "black-box" it is impossible to construct an explanation of the box's behavior that will support such predictions. A full, causal understanding of behavior requires knowledge of both the information processing mechanisms and the input data.

It is interesting to note that the persistent asymmetry found throughout the simulations and experiments reported in this article are typical of a well-studied phenomenon in connectionist modeling: catastrophic interference. Catastrophic interference refers to the fact that subsequent learning can overwrite prior learning in connectionist networks (see French, 1999, for a complete review). Depending on the similarity between item A and item B, the subsequent learning of B after A may completely erase any memory of A; hence, the name catastrophic interference. In connectionist networks, catastrophic interference arises because of the overlap in the internal representations developed across the hidden units for similar input stimuli (French, 1992).

Interpreting the asymmetry effects discussed throughout this article as examples of catastrophic interference (in an associative neural network memory model) allows us to extend the scope of this model to account for an illusive interference effect in early infant visual memory. During the mid- to late-seventies there was a debate surrounding the robustness of infant visual memory. A number of labs (e.g., Deloache, 1976; Fagan, 1973; McCall, Kennedy, & Dodds, 1977) suggested that infants suffer from substantial forgetting if presented with new material during the retention interval. These studies relied on a habituation procedure. Infants were habituated to a first image (A). After habituation to this image, they were habituated to a second image (B). After this second habituation phase, infants were presented with A again. A release in habituation to A (as measured by a renewed interest in A) was interpreted as suggesting that the intervening

habituation to B had caused the memory of A to disappear. The puzzling thing was that retroactive interference did not occur with all stimuli. In some cases interference occurred whereas in others it did not. Other investigations of loss in visual memory (using different stimuli) did not report any interference (Cohen, Deloache, & Pearle, 1977; Fagan 1977).

This debate was never completely resolved. However, two general conclusions were drawn. First, the similarity between the two images A and B seemed to have an impact on whether interference would occur or not but it was unclear how. Second, it was necessary for the infant to attend to and encode the features of the second stimulus (B) for interference to occur. We would like to suggest that performance on the categorization and memory tasks described above reflect the operation of the same information processing mechanisms. Namely, they reflect the way in which information is stored in an associative system with distributed representations typical of connectionist networks.

This interpretation underscores the intimate link between memory and categorization in preferential looking tasks. While memory and categorization have been studied together within the context of conditioned leg responses (Hayne, Rovee-Collier, & Perris, 1987) there have been no attempts to link visual recognition memory explicitly with preferential looking based categorization. The connectionist model provides a single mechanistic account of both memory and categorization and illustrates how considering issues in terms of information processing mechanisms can lead to the synthesis of apparently independent fields of study.

It is also important to understand the limitations of this model. All models are approximations. Building a model is not the same as building an infant. We do not wish to suggest that all of infant visual cognition can be accounted for with simple autoencoder networks. Nor do we wish to suggest that the infant is essentially an autoencoder. We have used the autoencoder to model looking time behaviors because autoencoders capture the representation construction hypothesis implicit in verbal descriptions of habituation (Charlesworth, 1969; Cohen, 1973; Solokov, 1963). The simulation work in this article suggests that autoencoders are part of the same class of learning systems as those used by the infants to learn perceptual categories (and therefore share their computational properties).

In summary, this paper has reported on a connectionist model of infant visual categorization. Asymmetric category exclusivity was found to arise from combination of connectionist information processing and the statistical distribution of features in the familiarization stimuli. The model predicted that asymmetric exclusivity would persist in the face of mixed familiarization category learning. An empirical studies with 3- to 4-month-olds confirmed the model predictions. Finally, a consideration of the mechanisms that underlie categorization underscores the intimate link that exists between visual memory and categorization.

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzman machines. Cognitive Science, 9, 147-169.

Bullinaria, J. (1995). Modeling reaction times. In L. Smith & P. Hancock (Eds.). Neural computation and psychology (pp. 34-48). London, UK: Springer.

Charlesworth, W. R. (1969). The role of surprise in cognitive development. In D. Elkind & J. Flavell (Eds.), Studies in cognitive development. Essays in honor of Jean Piaget (pp. 257-314). Oxford, UK: Oxford University Press.

Cohen, L. B. (1973). A two-process model of infant visual attention. Merrill-Palmer Quarterly, 19, 157-180.

Cohen, L. B., Deloache, J. S., & Pearl, R. A. (1977). An examination of interference effects in infants' memory for faces. Child Development, 48, 88-96.

Cottrell, G. W., Munro, P., & Zipser, D. (1988). Image compression by backpropagation: an example of extensional programming. In N. E. Sharkey (Ed.), Advances in cognitive science, Vol. 3 (pp. 208-240). Norwood, NJ: Ablex.

Deloache, J. S. (1976). Rate of habituation and visual memory in infants. Child Development, 47, 145-154.

Di Lollo, V. & Enns, J. T. & Rensink, R. A. (in press). Competition for consciousness among visual events: The psychophysics of reentrant visual processes. Journal of Experimental Psychology: General.

Eimas, P. D., & Quinn, P. C. (1994). Studies on the formation of perceptually based basic-level categories in young infants. Child Development, 65, 903-917.

Eimas, P. D., Quinn, P. C., & Cowan, P. (1994). Development of exclusivity in perceptually based categories of young infants. Journal of Experimental Child Psychology, 58, 418-431.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). Rethinking innateness: A connectionist perspective on development. Cambridge, MA: MIT Press.

Fagan, J. F. III (1970). Memory in the infant. Journal of Experimental Child Psychology, 9, 217-226.

Fagan, J. F. III (1973). Infant delayed recognition memory and forgetting. Journal of Experimental Child Psychology, 16, 424-450.

Fagan, J. F. III (1977). Infant recognition memory: Studies in forgetting. Child Development, 48, 68-78.

Fahlman, S. E. (1988). Faster-learning variations on back-propagation: An empirical study. In D. S. Touretzky, G. E. Hinton, & T. J. Sejnowski (Eds.), Proceedings of the 1988 Connectionist Models Summer School (pp. 38-51). Los Altos, CA: Morgan Kaufmann.

Fantz, R. L. (1964). Visual experience in infants: Decreased attention to familiar patterns relative to novel ones. Science, 164, 668-670.

French, R. M. (1992). Semi-distributed representations and catastrophic forgetting in connectionist networks. Connection Science, 4, 365-377.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. Trends in Cognitive Science, 3, 128-135.

Haaf, R. A., Brewster, M., de Saint-Victor, C. M., & Smith, P. H. (1989). Observer accuracy and observer agreement in measurement of visual fixation with fixed trial procedures. Infant Behavior and Development, 12, 211-230.

Hayne, H., Rovee-Collier, C, & Perris, E. E. (1987). Categorization and memory retrieval by three-month-olds. Child Development, 58, 750-767.

Hertz, J., Krogh, A., & Palmer, R. G. (1991). Introduction to the theory of neural computation. Reading, MA: Addison Wesley.

Homa, D., & Chambliss, D. (1975). The relative contributions of common and distinctive information on the abstraction from ill-defined categories. Journal of Experimental Psychology: Human Learning and Memory, 1, 351-359.

Madole, K. L., & Oakes, L. M. (1999). Making sense of infant categorization: Stable processes and changing representations. Developmental Review, 19, 263-296.

Mareschal, D. (in press). Connectionist methods in infancy research. J. W. Fagen & H. Hayne (Eds.), Progress in infancy research, vol. 2. Mahwah, NJ: Erlbaum.

Mareschal, D., Plunkett, K., & Harris, P. (1999). A computational and neuropsychological account of object-oriented behaviours in infancy. Developmental Science, 2, 306-317.

Mareschal, D., & Shultz, T. R. (1996). Generative connectionist networks and constructivist cognitive development. Cognitive Development, 11, 571-603.

Mareschal, D., & French, R. M. (1997). A connectionist account of interference effects in early infant memory and categorization. In M. G. Shafto & P. Langley (Eds.), Proceedings of the 19th annual conference of the Cognitive Science Society(pp. 484-489). Mahwah, NJ: Erlbaum.

Mareschal, D. & French, R. M. (2000) Mechanisms of categorization in infancy. Infancy,1 , 59-76.

McCall, R. B., Kennedy, C. B., & Dodds, C. (1977). The interfering effect of distracting stimuli on infant's memory. Child Development, 48,79-87.

McLeod, P., Plunkett, K. and Rolls, E.T. (1998) Introduction to connectionist modeling of cognitive processes. Oxford: Oxford University Press.

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

Oakes, L. M.. Plumert, J. M., Lansink, J. M., & Merryman, J. D. (1996). Evidence for task-dependent categorization in infancy. Developmental Psychology, 19, 425-440.

O'Neill, J. M., Jacobson, S. W., & Jacobson, J. L. (1994). Evidence of observer reliability for the Fagan Test of Infant Intelligence (FTII). Infant Behavior and Development, 17, 465-469.

Plunkett, K. & Sinha, C. (1992). Connectionism and developmental theory. British Journal of Developmental Psychology, 10, 209-254.

Plunkett, K. & Elman, J. L. (1997). Exercises in rethinking innateness. Cambridge, MA: MIT press.

Quinn, P. C. (1987). The categorical representation of visual pattern information by young infants. Cognition, 27, 145-179.

Quinn, P. C. (1998). Object and spatial categorization in young infants: "What" and "where" in early visual perception. In A. M. Slater (Ed.), Perceptual development:

Visual, auditory, and speech perception in infancy (pp. 131-165). East Sussex, UK: Psychology Press (Taylor & Francis).

Quinn, P. C., & Eimas, P. D. (1996a). Perceptual organization and categorization in young infants. Advances in Infancy Research, 10, 1-36.

Quinn, P. C., & Eimas, P. D. (1996b). Perceptual cues that permit categorical differentiation of animal species by infants. Journal of Experimental Child Psychology, 63, 189-211.

Quinn, P. C., Eimas, P. D., & Rosenkrantz, S. L. (1993). Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants, Perception, 22, 463-475.

Quinn, P. C., & Johnson, M. H. (1997). The emergence of perceptual category representations in young infants. Journal of Experimental Child Psychology, 66, 236-263.

Rumelhart, D. & McClelland, J. (1986). Parallel Distributed Processing, Vol. 1. Cambridge, MA: The MIT Press.

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

Schafer, G. & Mareschal, D. (in press). Modeling infant speech sound discrimination using simple associative networks. Infancy, 1(3).

Schuler, E. M. (1980). Simon and Schuster's guide to dogs. New York: Simon and Schuster.

Siegal, M. (1983). Simon and Schuster's guide to cats. New York: Simon and Schuster.

Sokolov, E. N. (1963). Perception and the conditioned reflex. Hillsdale, NJ: Erlbaum.

Slater, A. M. (1995). Visual perception and memory at birth. In C. Rovee-Collier & L. P. Lipsitt (Eds.), Advances in Infancy Research (Vol. 9, pp. 107-162). Norwood, NJ: Ablex.

Younger, B, A. (1985). The segregation of items into categories by ten-month-old infants. Child Development, 56, 1574-1583.

Younger, B. A., & Cohen, L. B. (1986). Developmental changes in infants' perception of correlation among attributes. Child Development, 57, 803-815.

Younger, B. A., & Fearing, D. D. (1999). Parsing items into separate categories: Developmental change in infant categorization. Child Development, 70, 291-303.

Appendix

Measurements (in millimeters) made from the original Quinn, Eimas, and Rosenkrantz (1993) cat and dog photographs. These data were used to train and test the networks described throughout this article.

Measurements of Cat exemplars

	head length	head width	eye separation	ear separation	ear length	nose length	nose width	leg length	vertical extent	horizontal extent
cat1	29	32	7	28	12	0	3	0	54	62
cat2	12	13	4	12	5	3	2	14	25	50
cat3	20	20	4	17	6	5	3	15	26	67
cat4	13	17	4	17	5	3	2	28	28	46
cat5	13	14	4	14	4	4	3	15	23	42
cat6	18	22	3	17	6	6	3	24	42	70
cat7	10	12	3	7	3	2	1	24	24	47
cat8	23	24	5	26	7	4	4	25	50	64
cat9	16	17	4	15	5	5	4	22	32	54
cat10	16	15	3	12	8	3	2	15	30	65
cat11	19	27	5	20	8	4	3	22	71	57
cat12	19	21	4	12	5	5	4	20	39	65
cat13	25	30	6	30	14	6	5	0	50	81
cat14	16	20	3	16	13	5	3	26	29	59
cat15	17	27	5	22	5	3	3	28	40	43
cat16	18	21	4	20	6	4	4	35	55	43
cat17	23	22	5	24	7	6	4	35	52	56
cat18	20	22	5	23	7	5	4	28	34	54

Measurements of Dog exemplars

	head length	head width	eye separation	ear separation	ear length	nose length	nose width	leg length	vertical extent	horizontal extent
dog1	16	22	0	0	16	6	7	25	21	53
dog2	23	16	0	2	8	5	8	35	21	42
dog3	16	16	4	13	5	7	6	25	26	64
dog4	20	24	4	11	7	10	10	29	22	47
dog5	15	22	4	0	20	10	6	31	34	55
dog6	13	15	3	4	8	6	4	25	19	41
dog7	15	20	3	5	9	8	5	28	26	60
dog8	13	9	4	12	8	7	5	19	20	49
dog9	15	21	3	10	19	3	3	32	20	46
dog10	33	30	11	37	12	3	4	40	50	66
dog11	17	17	5	13	6	7	5	28	22	55
dog12	29	21	6	31	15	15	13	31	28	58
dog13	19	15	6	20	19	10	9	34	46	44
dog14	25	20	6	28	15	10	8	28	30	55
dog15	21	24	7	0	15	10	8	20	32	49
dog16	23	20	7	23	15	8	6	26	34	36
dog17	16	21	6	0	10	7	10	28	21	62
dog18	14	22	3	0	15	9	6	24	26	30

Footnotes

1 This is a common way of relating response times to network error scores (e.g., Seidenberg & McClelland, 1989; Mareschal, Plunkett, & Harris, 1999; Quinn and Johnson, 1997; but see Bullinaria, 1995, for possible counter-arguments).

2 The parameter values were as follows: learning rate = .2, momentum = .9, and Fahlman offset = .1. A description of the role of the learning-rate and momentum can be found in Plunkett and Elman (1997). The Fahlman offset is discussed in Fahlman (1988).

3 This transformation preserves the covariation between cues. This is important because both infants (Younger, 1985) and autoencoder networks (Mareschal & French, 2000) have been shown to use covariation information in establishing category boundaries.

4 For practical reasons, a maximum criterion of 2500 epochs (10 times the 250 epoch criterion) was used to terminate any simulations that failed to reach the 0.2 error criterion. This is analogous to the fact that, in practice, any study with infants has a fixed maximum duration.

5 Although the fitted normal distributions may predict negative feature value (e.g., eye separation) the actual values used to train networks were always between 0 and 1. The negative predicted values reflect the presence of a skew in the underlying distribution of actual values.

Table 1.Mean fixation times (in seconds) and standard deviations (in parentheses) during familiarization trials of Experiment 1.

Group	Trials 1-3	Trials 4-6
MAINLY-CAT	11.02 (2.56)	10.44 (2.99)
MAINLY-DOG	9.19 (3.32)	9.68 (3.61)

Table 2. Mean novel category preference for the two familiarization conditions of Experiment 1

	Familiarization condition	
	MAINLY-CAT	MAINLY-DOG
Mean Novel Category Pref.	58.32	47.73
SD	19.40	15.56
N	24	24
t(vs. chance)	2.10*	-0.71

Note: * $p<0.025$, one-tailed

Figure Captions

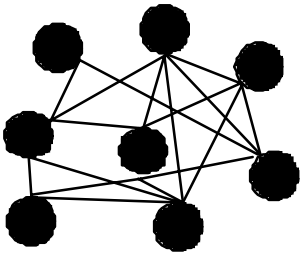
Figure 1. Schema of (a) a generic and (b) a feed-forward connectionist network.

Figure 2. Mean error prior to familiarization, after familiarization, and on novel exemplars after familiarization for networks trained with cats or dogs.

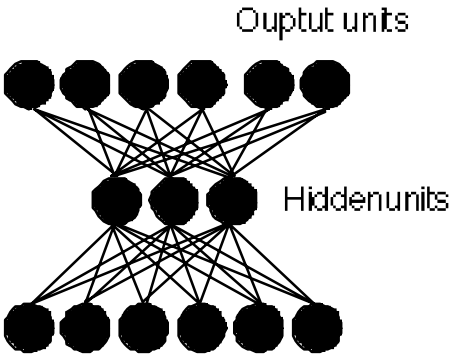
Figure 3. Mean error on a novel cat and a novel dog exemplar for networks trained on cats or dogs.

Figure 4. Gaussian probability distributions generated from the means and standard deviations of normalized cat features (thin line) and dog features (thick line). The area under each curve sums to 1.0.

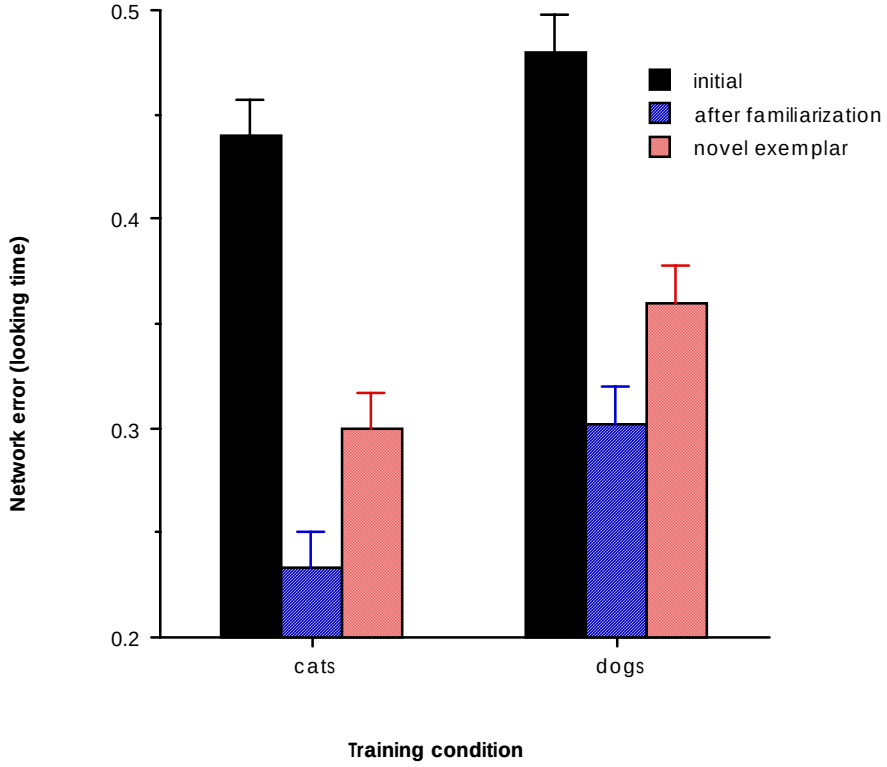
Figure 5. Mean error on a novel cat and a novel dog exemplar for networks trained with mainly cats (8 cats and 4 dogs) or mainly dogs (8 dogs and 4 cats).

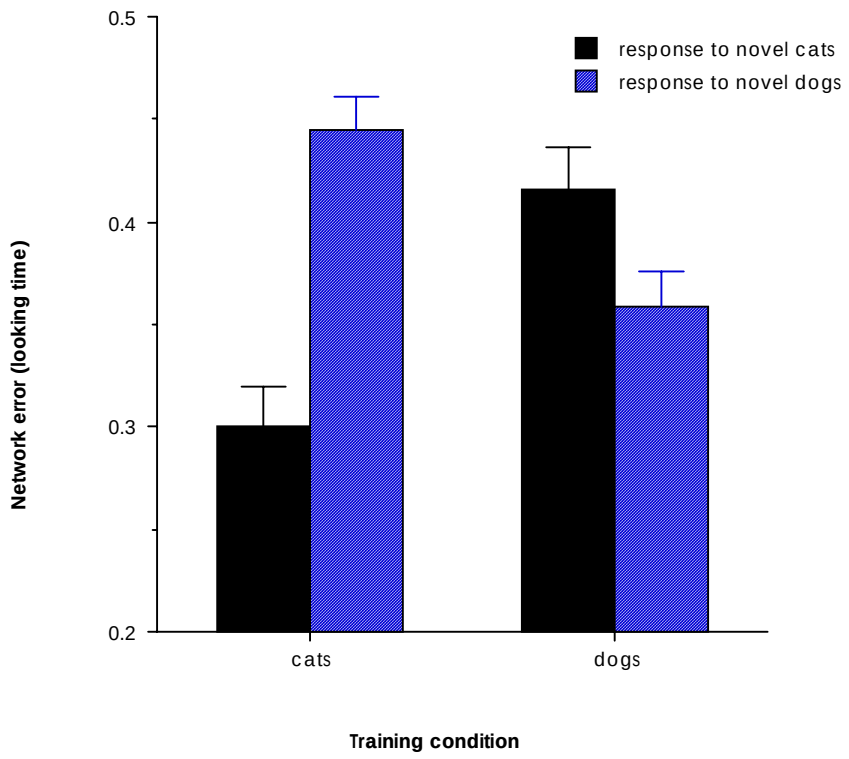


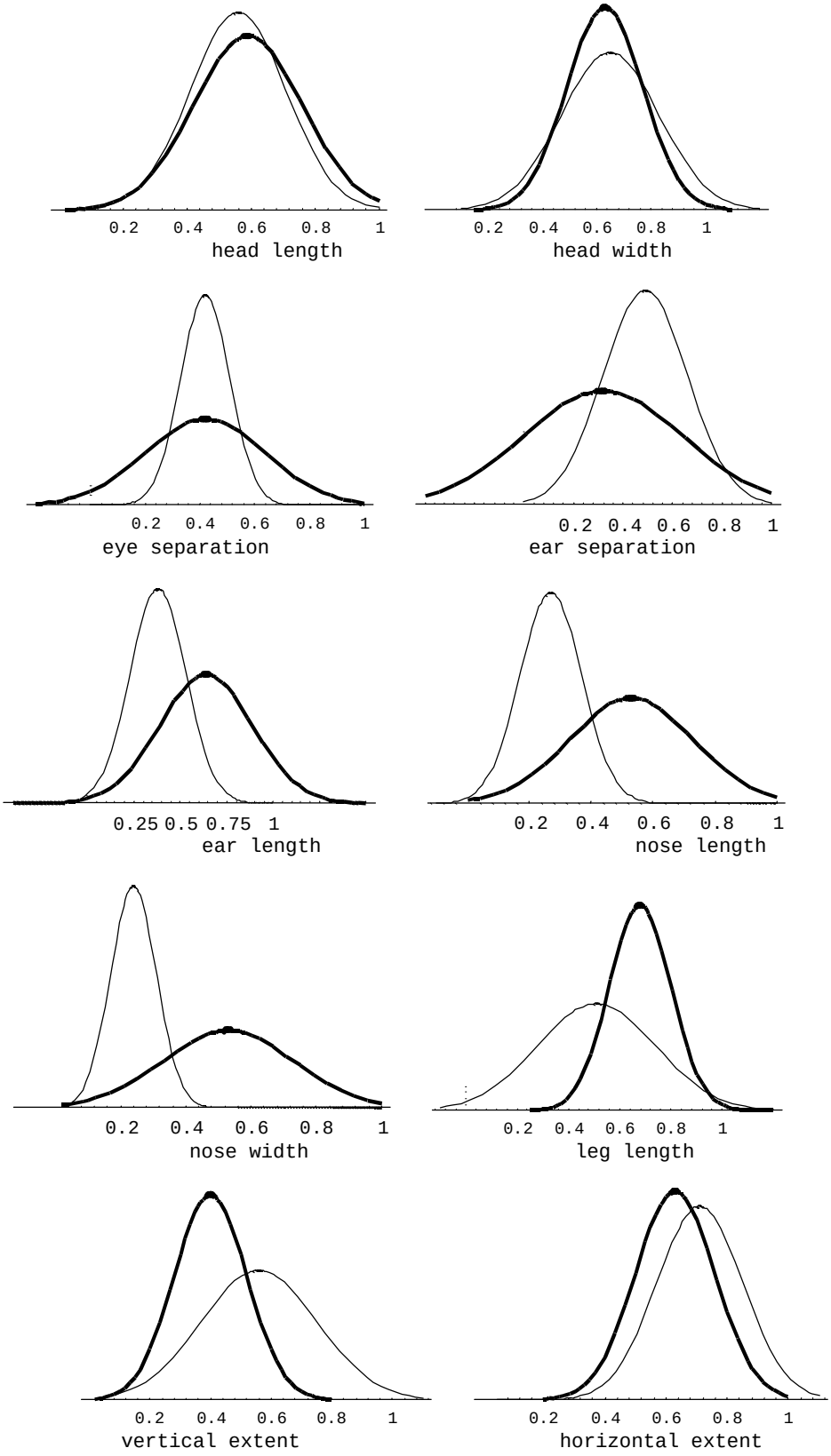
a.

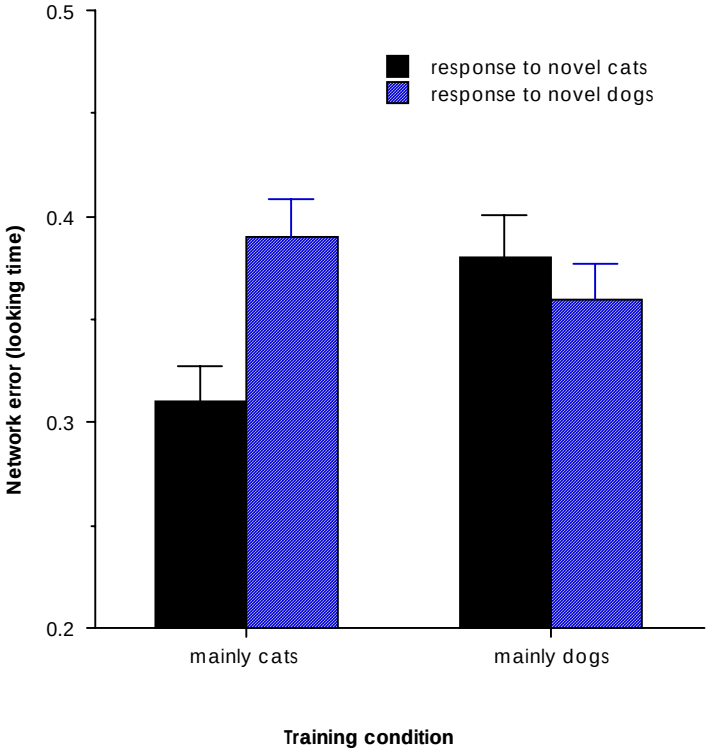


b.


0.5
0.4
0.3
0.2
Network error (looking time)
initial
after familiarization
novel exemplar
cats
dogs
Training condition


0.5
0.4
0.3
0.2
response to novel cats
response to novel dogs
Network error (looking time)
cats
dogs
Training condition


0.2 0.4 0.6 0.8 1
head length
0.2 0.4 0.6 0.8 1
head width
0.2 0.4 0.6 0.8 1
eye separation
0.2 0.4 0.6 0.8 1
ear separation
0.25 0.5 0.75 1
ear length
0.2 0.4 0.6 0.8 1
nose length
0.2 0.4 0.6 0.8 1
nose width
0.2 0.4 0.6 0.8 1
leg length
0.2 0.4 0.6 0.8 1
vertical extent
0.2 0.4 0.6 0.8 1
horizontal extent


0.5
0.4
0.3
0.2
Network error (looking time)
response to novel cats
response to novel dogs
mainly cats
mainly dogs
Training condition