ARTICLE WITH PEER COMMENTARIES AND RESPONSE

Learning to perceive object unity: a connectionist account

Denis Mareschal[1] and Scott P. Johnson[2]

1. *Centre for Brain and Cognitive Development, Birkbeck College, London, UK*
2. *Department of Psychology, Cornell University, USA*

Abstract

To explore questions of how human infants begin to perceive partly occluded objects, we devised two connectionist models of perceptual development. The models were endowed with an existing ability to detect several kinds of visual information that have been found important in infants' and adults' perception of object unity (motion, co-motion, common motion, relatability, parallelism, texture and T-junctions). They were then presented with stimuli consisting of either one or two objects and an occluding screen. The models' task was to determine whether the object or objects were joined when such a percept was ambiguous, after specified amounts of training with events in which a subset of possible visual information was provided. The model that was trained in an enriched environment achieved superior levels of performance and was able to generalize veridical percepts to a wide range of novel stimuli. Implications for perceptual development in humans, current theories of development and origins of knowledge are discussed.

Introduction

We inhabit a visual world that is filled with objects. Many of the objects we see are partly occluded by other, nearer surfaces, and it is routine for objects to go in and out of sight. Our impression of this visual array, nevertheless, is not one of fleeting or partial images (consistent with what is projected onto the retina), but rather an environment composed of solid, continuous, permanent entities. The visual system, therefore, is adept at imparting structure to an incompletely specified visual array. How does this way of experiencing the world arise? Does the young infant possess similar percepts to adults, in that he or she is born with impressions of segregated, coherent objects at various distances? Or does the infant's visual world consist of a series of disjoint, unrelated shapes that do not cohere into a sensible array until some period of development?

These questions have long interested philosophers and psychologists. James (1890) described the neonate's perceptual experience as fundamentally chaotic: 'The baby, assailed by eyes, ears, nose, skin, and entrails at once, feels it all as one great blooming, buzzing confusion' (vol. 1, p. 488). James went on to suggest that 'Infants must go through a long education of eye and ear before they can perceive the realities which adults perceive. *Every perception is an acquired perception*' (vol. 2, p. 78; emphasis in original). This position was echoed by Piaget (1952, 1954), who proposed that at birth, the infant's visual world consists of a patchwork or 'tableaux' of moving colors and shapes, as opposed to segregated, coherent objects. Perceptual organization was thought to emerge only gradually over the first two postnatal years, via direct manual experience with objects and coordination of visual, auditory and tactile information.

More recent work on infants' object perception has called into question these descriptions of young infants' capabilities and experiences. For example, Kellman and Spelke (1983) investigated the conditions under which 4-month-old infants perceive the unity of two surfaces (e.g. two rod parts) that extend from behind a nearer, occluding box (Figure 1a). Kellman and Spelke (1983; Kellman, Spelke & Short, 1987) found that after habituation to a display in which the two surfaces underwent common motion behind a stationary occluder (reported by adults to consist of a single object behind an occluder), the infants looked longer at two disjoint rod parts (a 'broken' rod; see Figure 1b) than at a single,

Address for correspondence: Denis Mareschal, Centre for Brain and Cognitive Development, School of Psychology, Birkbeck College, Malet Street, London WC1E 7HX, UK; e-mail: d.mareschal@bbk.ac.uk or Scott P. Johnson, Department of Psychology, Uris Hall, Cornell University, Ithaca, NY, 14853, USA; e-mail: sj75@cornell.edu

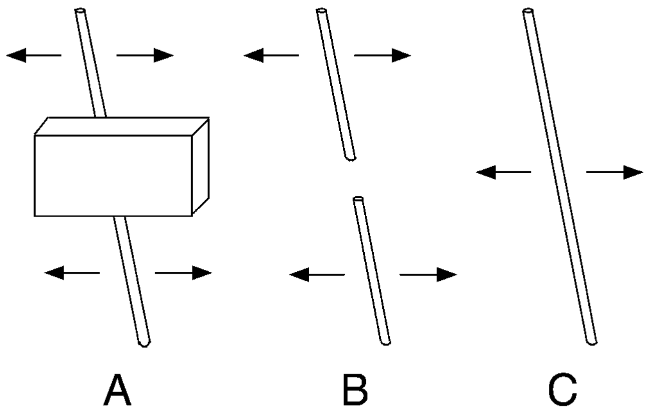


Figure 1 *Displays used by Kellman and Spelke (1983) to explore young infants' perception of object unity. A: A partly occluded rod moves relative to a stationary occluder. B: Broken rod. C: Complete rod. After habituation to A, infants often show a preference for B relative to C, indicating perception of the rod's unity in A.*

complete rod (Figure 1c). Given infants' tendency to look longer at novel relative to familiar stimuli after a period of habituation (Bornstein, 1985), this result suggests that these infants perceived the rod surfaces in the habituation display as occupying a single, partly occluded object. That is, the infants did not respond only to what was directly visible, but instead responded according to the distal characteristics of the event in the habituation display. This experiment and others (see Johnson, 2000; Needham, Baillargeon & Kaufman, 1997 for reviews), therefore, make clear that young infants' object perception skills are more sophisticated than allowed for by James or Piaget.

Two further conclusions were drawn from the Kellman and Spelke (1983) experiments. The first concerned the kinds of visual information employed early after birth in object perception tasks. Kellman and Spelke tested 4-month-olds' responses to a variety of displays that appeared to adults to consist of two objects, one partly occluded by another. In contrast to adults, the infants seemed to perceive object unity only when the partly hidden surfaces moved relative to their surroundings. Infants were posited not to take advantage of such potential visual information for object unity as the colors, textures and shapes of surfaces, relying solely on common motion. A 'two-process' account of unit formation has since been proposed (Kellman, 1996). For infants younger than 6 months, common motion of surfaces that lead behind an occluder is both necessary and sufficient to specify their unity. Only after 6 months do infants utilize additional sources of information for unity, such as surface appearance, and edge and surface orientation.

A second conclusion drawn from the early work on object unity concerned the possibility that for young infants, some percepts and concepts are qualitatively similar to those of adults: 'Humans may begin life with the notion that the environment is composed of things that are coherent, that move as units independently of one another, and that tend to persist, maintaining their coherence and boundaries as they move' (Kellman & Spelke, 1983, p. 521). Spelke (1990, 1994) has since proposed that the earliest kinds of object perception can be characterized as reasoning in accord with fundamental physical principles. One of these is the principle of contact: visible surfaces that undergo a common, rigid motion tend to be connected (Spelke & Van de Walle, 1993).

More recent research has explored further both the possibility that young infants utilize only a limited range of available visual information in object perception tasks, and the notion that core principles guide early object perception. In the next two sections of this paper, some of this evidence will be presented. Notably, there is not yet an adequate account of the development of object perception that can encompass the full range of evidence, although progress has been made toward such a theory. We describe in a subsequent section computational models that were designed to investigate whether and how the perception of object unity in an ambiguous stimulus (such as depicted in Figure 1a) might be learned. Before describing the models, we review evidence concerning the roles of various sources of information in young infants' perception of object unity, and the ontogenetic origins of this skill.

What visual cues are important in young infants' object perception?

Johnson and colleagues (Johnson & Aslin, 1996; Johnson & Náñez, 1995) probed in detail the kinds of visual information 4-month-olds use in object unity tasks. The first question was whether depth cues (binocular disparity, motion parallax, and accommodation and convergence, all potentially available in the Kellman & Spelke, 1983 rod-and-box displays) were necessary for perception of object unity in this age group (Johnson & Náñez, 1995). This was investigated with a two-dimensional, computer-generated display consisting of two rod parts, undergoing common motion, above and below an occluding box (Figure 2a). The objects were presented against a textured background consisting of a regular array of dots, in like manner to the Kellman and Spelke displays. After the infants were habituated to the rod-and-box display, they preferred a broken rod relative to a complete rod, replicating the Kellman and Spelke results. This implies that remaining information in the display was sufficient

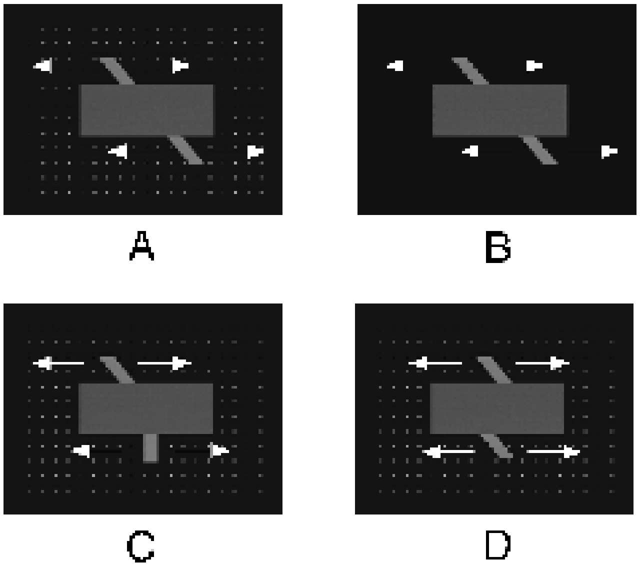


Figure 2 *Displays employed to investigate the role of texture and edge orientation in young infants' perception of object unity. A: Rod parts are aligned across the occluder, against a textured background. As the rod moves, it covers and uncovers progressively the texture, providing depth information. B: Rod parts are aligned across the occluder, against a matte black background, with no texture information for depth. C: Rod parts are not aligned, but are relatable (if extended, they would meet behind the occluder). D: Rod parts are neither aligned nor relatable. Four-month-old infants perceive unity only in A, underscoring the importance of edge alignment and texture to veridical object percepts. (Adapted from Johnson & Aslin, 1996.)*

to support perception of object unity in 4-month-olds (i.e. three-dimensional depth cues are not necessary for young infants to perceive unity).

Johnson and Aslin (1996) went on to vary systematically the cue availability in two-dimensional rod-and-box displays. In one experiment, they assessed 4-month-olds' perception of object unity in displays without background texture (Figure 2b), asking if accretion and deletion of background texture may have contributed to the process, perhaps as a depth cue. This cue was available in previous experiments with both three-dimensional (Kellman and Spelke, 1983) and two-dimensional (Johnson and Náñez, 1995) displays. Interestingly, the infants preferred neither the broken nor complete rod test display, implying no clear percept of unity during habituation. A recent study of 4-month-olds' perception of transparency in two-dimensional displays found that when background texture was visible both around and 'through' a surface, the infants responded as if the surface was translucent, but when texture was visible only around the surface (and not through it), the infants appeared to perceive it as opaque (although adults judged this latter stimulus to contain a translucent object; Johnson & Aslin, 2000). In two-dimensional displays, therefore, background texture may be necessary for segregation of visible surfaces into their constituent depth planes by this age group.

Johnson and Aslin (1996) next explored the role of orientation of the rod parts' edges, asking if misaligned edges may also impact perception of the rod's unity. This was accomplished in two ways. First, a display was constructed with rod edges that were not aligned, but were *relatable* – that is, the edges would meet at an angle greater than 90° if extended behind the occluder (Figure 2c; see Kellman & Shipley, 1991 for a formal definition of relatability). Second, a display was devised in which the rod edges were neither aligned nor relatable (Figure 2d). In both conditions, posthabituation test displays (broken and complete rods) matched the visible rod portions in the habituation display. In the former condition, there was no consistent test display preference, and in the latter condition, there was a preference for the complete rod. These two findings imply that the infants attended to rod orientation in perception of its unity: when edges are misaligned (Figure 2c), perception of object unity appears to be indeterminate, and when edges are neither aligned nor relatable, infants seem to perceive disjoint objects (Johnson, Bremner, Slater & Mason, 2000 and Smith, Johnson & Spelke, in press recently obtained similar results). Note that in all three of these conditions, the rod parts underwent common motion, and thus would be predicted to specify unity to 4-month-olds on the Kellman (1996) account of unit formation.

How does perception of object unity develop?

A second line of research has addressed the conclusion that humans begin postnatal life with certain kinds of object reasoning skills (Spelke, 1990, 1994; Spelke & Van de Walle, 1993). In an investigation of the possibility that perception of object unity is available from birth, Slater, Morison, Somers, Mattock, Brown and Taylor (1990) tested neonates with rod-and-box displays and reported consistently longer looking at a complete rod, relative to a broken rod, the opposite result relative to findings with 4-month-olds (Kellman & Spelke, 1983). This result demonstrates that neonates achieved figure –ground segregation in rod-and-box displays, clearly distinguishing the rod parts from the occluder and background, but they did not appear to perceive the unity of the rod parts. Instead, the neonates responded only to what was directly visible in the display, failing to make the 'perceptual inference' necessary to posit the existence of the hidden portion of the rod.[1]

This finding with neonates implies further that veridical perception of object unity, in the sense that performance corresponds to that of adults, emerges some time between birth and 4 months of age. This possibility was explored by habituating 2-month-olds with the rod-and-box display that had been shown previously to 4-month-olds (in which the older infants had apparently perceived the rod parts' unity), followed by the same complete and broken rod test displays (Johnson & Náñez, 1995). The younger infants showed no consistent posthabituation preference, suggesting that they had no clear percept of either unity or disjoint objects. It is possible, however, that the display presented to the 2-month-olds contained insufficient visual information to activate veridical surface segregation. This possibility was probed with displays in which this information was enhanced by showing more of the rod's surface (Johnson & Aslin, 1995). In this case, 2-month-olds preferred a broken rod display during test, indicating perception of the rod parts' unity during habituation. A similar logic was adopted in an investigation of neonates' perception of object unity in enhanced displays containing additional information relative to the displays used previously (by Slater *et al.*, 1990): more visible rod surface, greater depth difference, background texture, and so on (Slater, Johnson, Brown & Badenoch, 1996). Even with this additional information, however, the neonates preferred a complete rod during test, indicating perception of disjoint objects.

Progress toward a comprehensive account

The pattern of results across experiments leads to several conclusions. First, by 4 months, infants rely on multiple sources of information in object perception tasks: no single visual cue, such as common motion, drives perception of object unity. Second, perception of object unity *develops* – that is, surface segregation skills undergo change, improving rapidly after birth. There are no published reports of any other object perception task that has been presented to infants from birth through the first several postnatal months (see Johnson, 2000), and at present there is no direct evidence that would suggest that humans are born with object reasoning skills.

Despite this recent progress in our understanding of perceptual development, fundamental questions remain regarding the origins of object perception. We can make a start toward answering these questions by outlining some possibilities regarding perception of object unity. First, it might be that unity perception develops more or less as the visual system matures, and the infant is thereby able to take note of available information as improvements occur in acuity, color and luminance discrimination, depth perception, and so on. Second, infants may experience objects in accord with some core principles (such as contact), but may not exhibit evidence of these principles due to limitations in our testing procedures, or an inability to access the full range of available visual information that might trigger veridical percepts (see Jusczyk, Johnson, Spelke & Kennedy, 1999). Third, unity perception might be *learned*. On this account, visual skills are sufficient at birth (or very soon after birth) to abstract those visual cues specifying surface segregation, but the neonate fails to recognize that partly occluded and fully visible objects seen at different times might be one and the same. That is, visual sensitivity is sufficient to impart clear percepts of all visible surfaces in an array, but what is missing is the ability to link separated edges across a spatial gap.

What kind of evidence would allow us to distinguish between these contrasting views? One important tool with which to explore this and related questions is connectionist (computational) modeling, which has been successful in exploring a range of developmental phenomena (Elman, Bates, Johnson, Karmiloff-Smith, Parisi & Plunkett, 1996; Mareschal & Shultz, 1996; Mareschal, 2001). Connectionist models consist of networks of interconnected processing nodes, analogous to neurons, designed to learn through interactions with a specific 'environment' created by the modeler. Such models are often produced by arranging the nodes in layers, with connections within and between the layers. One common approach is the incorporation of an *input* layer that is responsible for initial processing of stimulus information, an *output* layer that provides a response, and an intermediary *hidden* layer that enables the internal 're-representation' of information in the environment. Representations are embodied in the weights assigned to connections or as patterns of activation across a bank of nodes, and are developed by extracting the statistical regularities present in the environment.

Computational models can provide rigorous and tangible accounts of development, because the time course and nature of learning can be captured and made explicit, and in implementing a model, the modeler is forced to make explicit what is meant by 'representation', 'acquired knowledge', 'innate knowledge', and so on. If a model can be shown to acquire a particular behavior, then the constraints built into that model (in terms of prior 'knowledge', information processing mechanisms and learning algorithms) constitute possible

[1] Note that both the neonate's and the adult's percepts are entirely consistent with the available evidence. What has changed is the bias in how the infants respond to ambiguous events.

candidate constraints on human learning. Of course, it is always possible that humans operate using a different set of constraints. Models do not provide definitive answers to questions of human information processing. What they provide, instead, is a set of possible solutions.

In the present article we report on two models of the development of perception of object unity, with the goal of explaining human performance across development. The models were first trained by exposure to simple events in a simulated, schematic visual environment. In these events, unified and disjoint 'objects' moved past and behind an occluding 'screen'. The models were endowed with the ability to extract object motion, common motion of two objects, accretion and deletion of texture, T-junctions (an intersection where one surface occludes another, so-called because the projected edge of the far surface stops at the edge of the near surface, analogous to the stem and bar of a T, respectively) and edge alignment. The models also possessed a short-term memory, such that when an object became occluded, a rapidly decaying trace of that object's representation remained. After varying amounts of training, we presented novel test events that incorporated the visual cues to which the model was sensitive. Test events always presented partly occluded (never fully visible) objects. Using the Johnson and Aslin (1996) and Kellman and Spelke (1983) strategy, we systematically included or omitted cues across displays and observed the models' responses. We found that after sufficient training, the models responded appropriately to object unity under conditions of partial occlusion, demonstrating the potential importance of learning in the development of object perception. The nature of the training environment (i.e. the cues that were made available) was critical in determining performance.

Building the models

The model assumptions

Our models were based on three assumptions: first, infants are capable of detecting visual information (such as common motion) prior to its effective utilization in object perception tasks; second, infants acquire the ability to perceive object unity in ambiguous stimuli through experience with a visual environment in which moving objects become occluded and unoccluded; and third, young infants are equipped with short-term memories. (Each assumption is discussed subsequently in more detail.) During a period of training, the models learned to associate certain perceptual cues with the presence of a single unified object or the presence of two disjoint objects. After training, the models applied this 'knowledge' to novel events in which object unity was not directly visible (i.e. when two rod parts were visible above and below the occluder). The key to learning to perceive unity in these ambiguous stimuli was the presence of a short-term perceptual memory and exposure to objects that *became* occluded and unoccluded.

The first assumption: detection, then utilization

Infants are born with a functional visual system, and exhibit marked preferences for some classes of stimuli over others: moving stimuli are preferred to static stimuli, patterned stimuli to unpatterned stimuli, high contrast to low contrast, and horizontal contours to vertical contours, among others (see Slater, 1995 for review). At birth, infants also provide evidence of figure–ground segregation. Recall that neonates preferred a complete to a broken rod after habituation to a rod-and-box display, implying perception of disjoint objects (i.e. two rod parts) in the original display (Slater *et al.*, 1996). When habituated to a complete rod in front of an occluder, in contrast, neonates subsequently preferred a broken rod test display (Slater *et al.*, 1990). These findings suggest that the neonates formed a clear impression of distinct, segregated surfaces in both displays: two rod parts separate from the occluder and background in the former condition, and a single rod separate from the occluder and background in the latter condition. Despite these visual skills, effective utilization of visual information in object segregation tasks lags behind its detection at birth. For example, T-junctions were available as cues for relative depth in the Slater *et al.* (1990, 1996) occluded-rod displays, but the neonates did not appear to have perceived occlusion in these displays. That is, the rod surfaces appear to have been perceived to *end* at the rod–box intersection, rather than continue behind, suggesting that T-junctions were detected (and contributed to figure–ground segregation) but misclassified as indicating edge *termination* rather than edge *continuation*. It is unlikely that the infants simply perceived the rod and occluder surfaces to occupy the same depth plane, because even at birth infants can distinguish objects at different distances (Slater, Mattock & Brown, 1990).

The potential role of surface motion in neonates' perceptual segregation is less clear, due to a complex developmental trajectory for motion sensitivity (see Banton & Bertenthal, 1997). It has been claimed that infants younger than 4 to 6 weeks of age lack cortical mechanisms subserving motion discrimination, a claim based in part on infants' preferential looking toward one side of a stimulus containing regions moving in opposite directions vs a uniform pattern on the other side

(Wattam-Bell, 1991, 1996a). In contrast to the Wattam-Bell experiments, evidence suggesting early motion sensitivity was obtained by LaPlante, Orr, Neville, Vorkapich and Sasso (1996) and Laplante, Orr, Vorkapich and Neville (2000) who demonstrated discrimination of translational and rotational direction in neonates, and Náñez (1988), who reported avoidance responses to looming stimuli in 3-week-olds. These apparently conflicting findings may be reconciled by considering that evidence of motion sensitivity in a particular experimental paradigm is strongly dependent on stimulus characteristics (e.g. slow vs fast velocity), type of motion (e.g. rotation vs looming) and methodology (e.g. paired vs sequential presentation) (see Banton & Bertenthal, 1997). It appears likely, then, that motion sensitivity is present prior to its effective contribution to all perceptual segregation tasks (cf. Wattam-Bell, 1996b for discussion).

The second assumption: learning object properties derives from visual experience

Neonates exhibit marked preferences for faces over other visual stimuli (Slater, 1995) and quickly learn to distinguish among the faces they see. For example, there is evidence that individual infants develop a preference for their own mother's face, presumably the most often-viewed face in the visual environment, within a few hours of birth (Walton, Bower & Bower, 1992; cf. Walton & Bower, 1993). This finding implies highly efficient mechanisms that are functional at birth to rapidly detect and discriminate salient visual stimuli, and to subsequently learn characteristics of these stimuli (Slater *et al.*, 1998; Slater, Bremner, Johnson, Sherwood, Hayes & Brown, 2000). Moreover, neonates have been shown to process compound stimuli as composed of a combination of attributes, indicating that at birth, infants attend to multiple aspects of individual displays (LaPlante *et al.*, 2000; Slater, Brown & Badenoch, 1997).

The third assumption: objects are remembered over a short interval

Neonates, like older infants, will habituate to repeated presentation of a stimulus, and recover interest to a novel stimulus (Slater, 1995). This suggests a functional short-term memory that guides neonates' attention to familiar and unfamiliar stimuli, and retains stimulus characteristics over brief intervals.

The model architecture

Figure 3 illustrates the model architecture. The models received input via a simple 'retina'. The information presented to the retina represented objects, their orientation and motions, and the background. This information was processed by seven encapsulated perceptual modules, each of which identified the presence of one of the following cues during specific portions of training and test events: (a) motion anywhere on the display; (b) co-motion of objects in the upper and lower halves of the display, whether in-phase or out-of-phase; (c) common motion of objects in the upper and lower halves of the display; (d) parallelism of object edges in the upper and lower halves of the display; (e) relatability of object edges in the upper and lower halves of the display; (f) texture deletion and accretion and (g) T-junctions. We chose these particular cues because of the importance of motion (i.e. cues a, b and c), edge orientation (cues d and e), and depth (cues f and g) to young infants' perception of object unity (Johnson & Aslin, 1996; Kellman & Spelke, 1983).

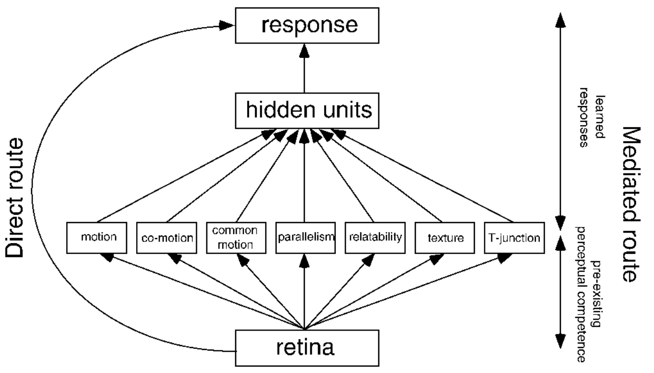


Figure 3 *Architecture of the models. See text for details.*

Each perceptual module fed into a layer of hidden units with sigmoid activation functions, which in turn fed into a response (output) layer. The response units determined the model's decision as to whether the ambiguous stimulus (i.e. the partly occluded rod) contained a single object, two disjoint objects, or neither (a response we termed 'indeterminate'). Unity was also a 'primitive', like the other cues, in that a model could perceive it directly in unambiguous cases (i.e. when the object was visible to one side of the occluder). These types of response to unity are consistent with evidence from human neonates. In the absence of any occlusion, neonates can discriminate between a broken and an unbroken visible rod. Indeed, this is a necessary precondition for interpreting the looking-time behaviors of neonates in experimental studies of the perception of object unity (e.g. Slater *et al.*, 1990). In the absence of direct perception (i.e. when the objects were partly occluded) the perception of unity was mediated by its association with other, directly perceivable, cues.

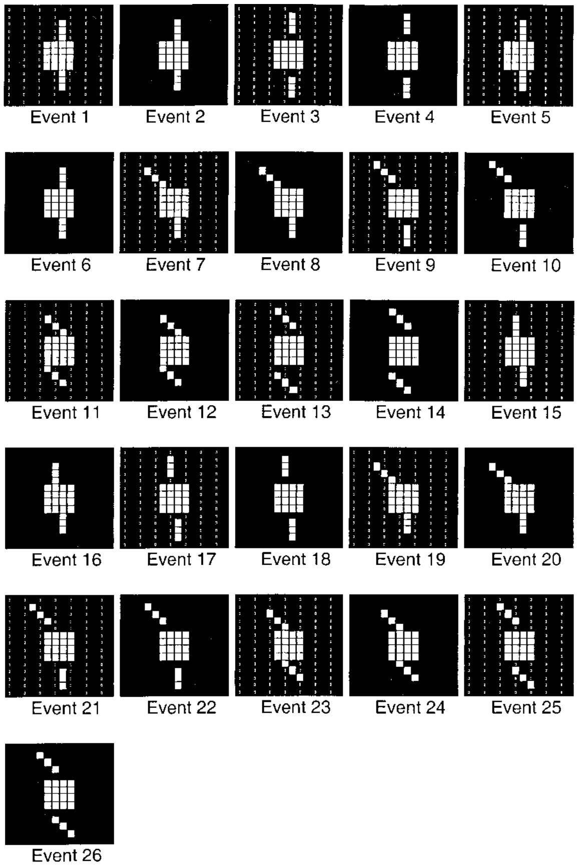


Figure 4 *Schematic depictions of training and test events. Refer to Figure 6 for the cues available in each event.*

Input and output representations and training

The models experienced input consisting of a central occluding screen, and either one long or two shorter objects (see Figure 4). The model's task was to determine whether one or two objects (not including the occluder) were present in the display. Feedback was provided by direct perception when the objects were not partly occluded (i.e. on either side of the screen) and by a decaying memory trace when direct perception was not possible. The models were tested for unity responses periodically after different training intervals, with displays in which object unity was potentially ambiguous. That is, testing was conducted with displays in which the objects were not visible on either side of the screen, only above and below it. (Testing of the models differed in an important respect relative to infants, because infants are observed for a posthabituation novelty preference. It seemed unnecessary to build a novelty preference into the models, given that we wanted to know if unity was perceived in ambiguous events, and this information was obtainable directly from the model.) Training and test displays were varied to incorporate or omit cues known to mediate perception of object unity: motion, alignment, relatability, T-junctions, and accretion and deletion of texture. Figure 6 lists the perceptual cues present in each event.

All events began with the object (or objects) moving on to the display from the side. During this initial portion of the event, the object was unobstructed from view. The object moved across the display, passed behind the area occupied by the occluding screen, reappeared on the other side of the screen, and continued off the display (Figure 5).

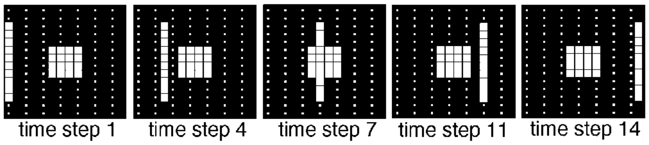


Figure 5 *Five time steps through training Event 1. The object (unified, in this case) appears both as fully visible and partly occluded at different times during the event.*

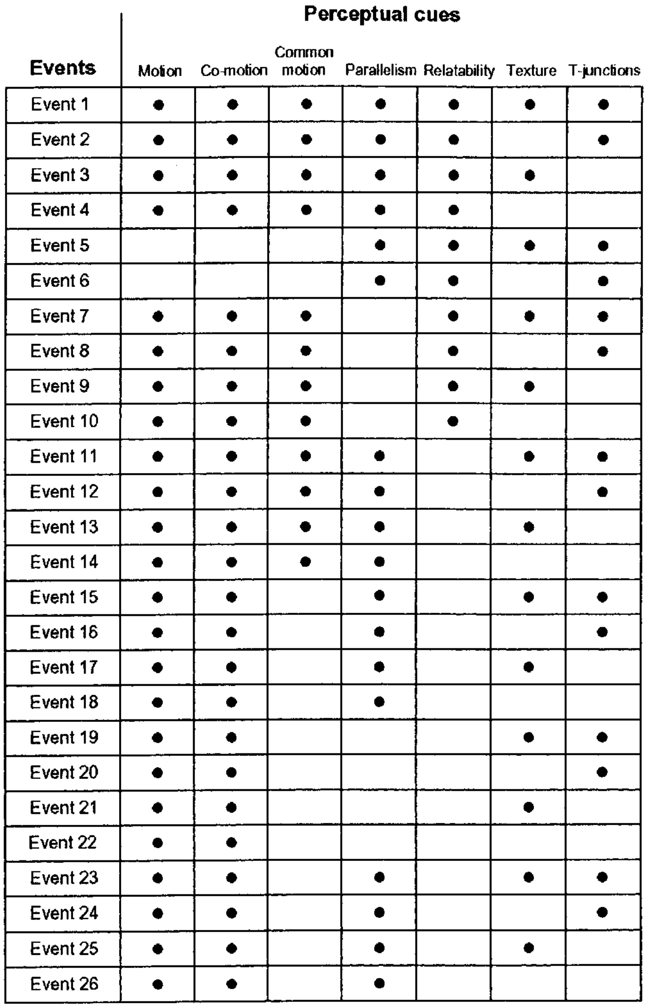
The perceptual modules

The bottom half of the network (see Figure 3) encompasses perceptual abilities that are functional at the earliest time of testing unity. Each module was designed to compute the presence or absence of a single cue in the displays experienced by the model. The modules incorporated general neural computational principles of summation, excitation, inhibition and local computation. However, there was no learning involved. The modules were tailored to the specific nature of the model's experience and were intended as analogues of the neonate's visual system, but not to embody its anatomy or physiology.[2] However, they did instantiate some of the basic principles believed to underlie the computation of the associated visual cues (see Spillman & Werner, 1990 for a review). For a more complete description of each perceptual module, see the Appendix.

What drives learning?

Learning was driven by an error feedback signal partly obtained directly from the environment and partly from

[2] The fact that these modules are operational at the earliest time of testing does not necessarily imply that they are 'hardwired' from birth. For example, Nakissa and Plunkett (1998) described a set of simulations in which networks evolve over many generations to become excellent learners of phonological discriminations. Networks from the final generation are unable to discriminate phonemes prior to any experience, but only a few minutes of real world speech are necessary for categorical perception of phonemes. One can imagine a similar scheme in which a short time of visual experience would fine-tune a set of crude perceptual modules, but at present this remains an empirical question.



Events	Perceptual cues						
	Motion	Co-motion	Common motion	Parallelism	Relatability	Texture	T-junctions
Event 1	●	●	●	●	●	●	●
Event 2	●	●	●	●	●		●
Event 3	●	●	●	●	●	●	
Event 4	●	●	●	●	●		
Event 5				●	●	●	●
Event 6				●	●		●
Event 7	●	●	●		●	●	●
Event 8	●	●	●		●		●
Event 9	●	●	●		●	●	
Event 10	●	●	●		●		
Event 11	●	●	●	●		●	●
Event 12	●	●	●	●			●
Event 13	●	●	●	●		●	
Event 14	●	●	●	●			
Event 15	●	●		●		●	●
Event 16	●	●		●			●
Event 17	●	●		●		●	
Event 18	●	●		●			
Event 19	●	●				●	●
Event 20	●	●					●
Event 21	●	●				●	
Event 22	●	●					
Event 23	●	●		●		●	●
Event 24	●	●		●			●
Event 25	●	●		●		●	
Event 26	●	●		●			

Figure 6 *Cues available in each event.*

memory. When the object was visible, the environment provided immediate feedback about the unity of the object via the direct perception link (see Figure 3). When the object was not completely visible, the environment could not provide feedback about the unity of the object. In this case, the model relied on a short-term, rapidly decaying memory. This memory was always active and encodes information obtained from direct perception (specifically unity information). Immediately following occlusion, the memory provided a trace of the state of the object prior to occlusion. After a short delay, that information decayed and was no longer available for learning about the relations between available cues and unity.

The relation between direct perception and memory is embodied in the target signal, T(t), used for training the network weights:

$$T_i(t) = E_i(t) + \mu.T_i(t - 1) \qquad (1)$$

with $-1 < T_i < +1$, $0 < \mu < 1$, and $E_i = 0.0$ when the rod is occluded. Ei is the unity feedback signal obtained from the environment (by direct perception) for output i, and μ is a parameter controlling the depth of memory. When $E_i = 0.0$ (i.e. there is no direct percept of unity), the target ($T_i(t)$) is derived entirely from the memory component $\mu.T_i(t - 1)$, the second term in the right-hand side of equation 1.

The weight updates are computed according to an error reduction algorithm (backpropagation) that minimizes the difference between the actual output and the target output activations. The system self-organizes in such a way as to minimize the difference between its unity prediction and what it perceives as true in its environment. There is no external agent providing the network with the desired answer. All target information required for updating weights is obtained directly from the environment (through direct perception) in the same way as the perceptual input is obtained, or from within the system (through memory). In other words, this is an example of *unsupervised* learning. Similar accounts of self-organization using backpropagation networks can be found elsewhere (e.g. Mareschal, French & Quinn, 2000; Munakata, McClelland, Johnson & Siegler, 1997; see also Baldi, Chauvin & Hornik, 1995, for formal proofs of the equivalence of some linear backpropagation networks with some linear self-organizing systems).

The model's unity response was driven by a combination of activation from the direct and mediated routes. When direct perception was possible, the activation from this route overrode that of the mediated route by saturating an output unit's response towards +1 or −1. When direct perception was not possible, the unity response was mediated through its associations with other cues that are directly available. In the present paper, we are interested in assessing the mediated route's performance. The degree to which the model's mediated response was correct when direct perception was *not* possible reflects how well it responded to incomplete information. The degree to which the mediated route's prediction was correct when direct perception *was* possible reflects how well the network has internalized general information about objects that applies across its entire learning environment. Network performance can be assessed either when direct perception is possible (events 3, 4 and 9 to 26 in Figure 4), or when it is not possible (e.g. on events 1, 2, 5, 6, 7 and 8).

In assessing the model's performance we compared the output of the mediated route with direct perception when available. When direct perception was not possible, the network's response was compared to the modeler's knowledge of what condition the ambiguous stimulus was derived from. A mediated response was scored as correct if it accurately predicted the origins of the ambiguous event (e.g. a unified object was perceived when the event was caused by a unified object). It was scored as incorrect when it predicted the opposite origins of the ambiguous event (e.g. two objects were perceived when a single object caused the event), and it was scored as indeterminate when the output was either (+1, +1) or (−1, −1). Because output units were linear, for the purposes of scoring the network responses the output values were classified as +1 if they were positive and −1 if they were negative. These responses were then compared with human responses under similar conditions to evaluate how well the model matches human data. A network's performance was tested by presenting it with events consisting of the ambiguous segment of the trajectory only. In other words, during testing, the networks could not use information available in the unambiguous segment of the trajectory to derive unity. The test results reported below were scored on what would correspond to time steps 7 and 8 of a full 14 time step event.

We report on two models that each contained the same architecture and training procedures previously described. Model 1 was trained in a 'simple' perceptual environment (a small subset of the events depicted in Figure 4), and Model 2 was trained with an 'enriched' environment (a larger subset of the events). To anticipate, we found that both models learned to predict unity in an ambiguous event, but the model that experienced an enriched environment acquired the most general knowledge of the relation between the presence of individual perceptual cues and the percept of unity.

Model 1: Learning in a simple environment

In the first model, ten networks were exposed to a world with minimal, but ecologically valid, constraints. These constraints correspond to events that are observable in a natural environment (and are, therefore, events with which even very young infants might have experience). The learning environment consisted of a single unified object moving across the display with or without background texture (events 1 and 2), and two co-linear disjoint objects moving across the display with or without texture (events 3 and 4). Learning occurred across the complete object trajectory (i.e. across both the ambiguous and unambiguous segments of the trajectory). Ten networks were tested, rather than only one, to explore differences in performance as a function of the starting points for training (i.e. random initial weights and events, described in the next paragraph). Events were randomly selected and presented to the networks one at a time. Each network, therefore, experienced its own idiosyncratic series of events determined by the random selection procedure. Networks were periodically tested for prediction of unity after specific intervals during training (10, 50, 100, 500, 1,000, 1,500, 2,000, 3,000, 4,000, 5,000, 6,000, 7,000 and 8,000 epochs).

Figure 7 shows the performance of the ten networks across all 26 events during testing. Consider first the four events on which Model 1 was trained (events 1 to 4). By 50 epochs, the networks correctly signaled the presence of a single object or two distinct objects on the *unambiguous* portion of the trajectory of these events (when the object elements are to the left or right of the occluder, not shown in Figure 7). When the networks were tested with the *ambiguous* portion of the trajectory (when the object elements are above and below the occluder), they quickly reached high levels of performance in all of the familiarization events. Learning was more rapid in events 3 and 4 because there are no ambiguous portions in these events. When tested with ambiguous events, accurate performance was delayed. Nevertheless, by 1,500 epochs all networks perceived events 1 and 2 as arising from a single, partly occluded object.

To understand the generality of knowledge encoded in the mediated route, consider next performance on events with which the networks were not trained (events 5 to 26, Figure 7). By 1,500 epochs, the networks performed correctly on the ambiguous portions of 15 of the remaining 22 test stimuli (68.2% of the events), but failed on events 10, 11, 19 to 21, 23 and 24. Note that events 15 to 26 are all displays with two objects in which there is no common motion. Failure on these events is tantamount to perceiving a single unified object even though the two object components are moving in the opposite direction. These networks also failed to perform correctly on 3 of the 26 events (events 9, 10 and 22) when tested with the visible unambiguous segment of the event, even after maximum training (8,000 epochs).

In summary, the networks that were trained with a simple perceptual environment learned to perceive object unity quite rapidly in most of the displays with unified objects, and perceived disjoint objects in most of the other displays. However, the networks were able to generalize their knowledge only to events that were relatively similar to those that were experienced during learning. It is not immediately clear why the networks failed to use some of the cue combinations appropriately. Inspection of conditions that led to unsuccessful performance does not lead to straightforward interpretation of all instances of failure, although it is notable that events 19, 20, 21, 23 and 24 each contain some combination of T-junctions, relatability and co-motion (but not common motion). These are all cues that can lead to the percept of unity, especially on the training set

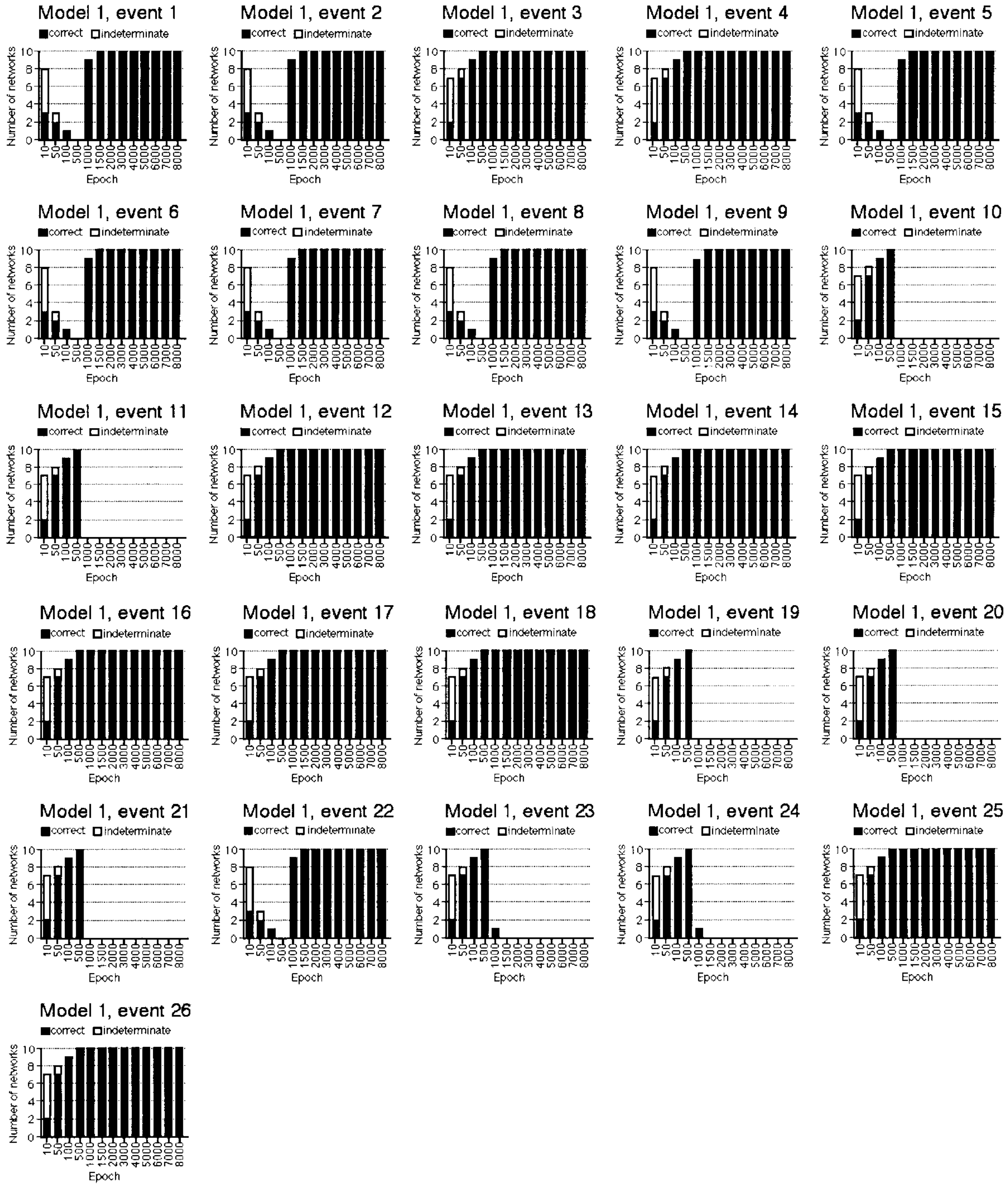
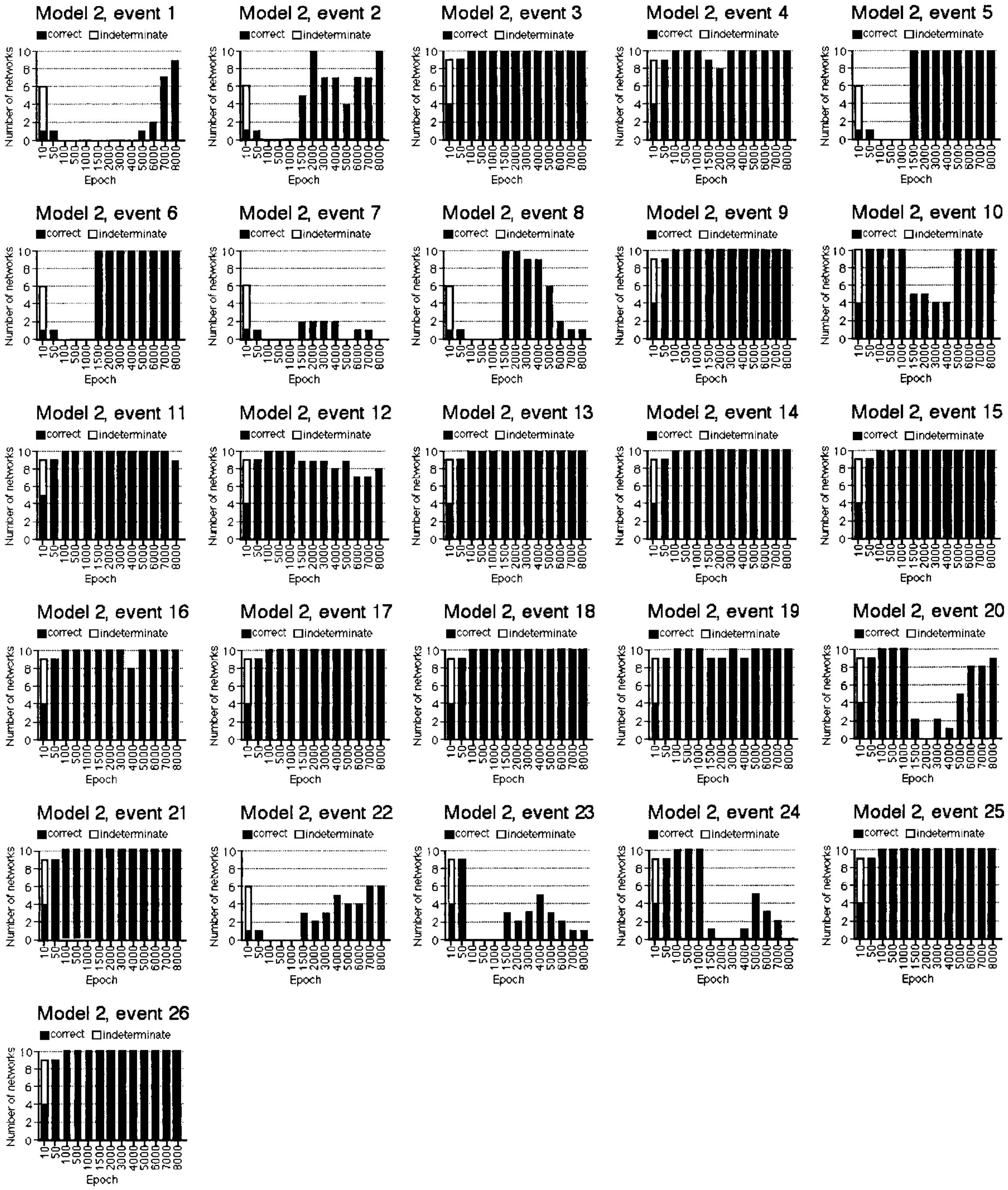


Figure 7 *Model 1 performance. Ten networks, each with random initial weights between nodes (perceptual modules, hidden units and output units), were trained with events 1–4. The networks were able to generalize learning to novel events, but performance was constrained by the limited training experience.*

experienced by the networks. Overall, then, performance was accurate across the ambiguous and unambiguous portions of the 26 events, but the few idiosyncracies suggest that impairments in performance might be tied to the limited set of training events.

Model 2: Learning in an enriched environment

In the second model, we extended the range of learning experiences by using a training set that was more representative of the entire range of events. This training set was



Model 2, event 1

Model 2, event 2

Model 2, event 3

Model 2, event 4

Model 2, event 5

Model 2, event 6

Model 2, event 7

Model 2, event 8

Model 2, event 9

Model 2, event 10

Model 2, event 11

Model 2, event 12

Model 2, event 13

Model 2, event 14

Model 2, event 15

Model 2, event 16

Model 2, event 17

Model 2, event 18

Model 2, event 19

Model 2, event 20

Model 2, event 21

Model 2, event 22

Model 2, event 23

Model 2, event 24

Model 2, event 25

Model 2, event 26

Figure 8 *Model 2 performance. Ten networks, each with random initial weights between nodes (perceptual modules, hidden units and output units), were trained with events 1, 2 and 17–22. The architecture was identical to the networks in Model 1, but performance was superior due to the enriched training environment.*

events 1, 2 and 17 to 22, providing Model 2 with additional exposure to disjoint objects relative to Model 1. Other aspects of the design, training and testing of the two models were identical. As was the case for Model 1, the mediated route in the Model 2 networks learned quickly (by 10 epochs) to detect either one or two objects in the unambiguous portion of the test events on which they were trained. The more rapid adaptation of the mediated route, relative to Model 1, was due to the more frequent exposure to disjoint objects in the training environment.

Figure 8 shows the performance of the ten networks during the ambiguous portions of all 26 events across testing. Consider first the events on which Model 2 was trained. By the end of training, the networks achieved high levels of performance in 7 of the 8 familiarization events (events 1, 2 and 17 to 21), and half the networks responded appropriately in the eighth (event 22). Initially, the networks perceived the single object in event 1 as arising from two disjoint objects, a tendency that was not overcome until after 4,000 epochs. There was then a period (from about 5,000 to 7,000 epochs) in which an increasing proportion of the population of networks perceived this stimulus as arising from a single unbroken object. There was much variation across the networks, therefore, as to how this event was perceived. The variation arose from the sequence of events experienced by the networks during training (i.e. a preponderance of disjoint objects), and the initial random weights prior to training. By 8,000 epochs, 9 of the 10 networks perceived this ambiguous event as arising from a single unbroken object, indicating that they had learned to go beyond a general default response that was consistent with the majority of training events (recall that the majority of training events arise from two distinct objects). A slightly different pattern emerged when tested for unity perception in event 2. As in event 1, the networks perceived disjoint objects during the initial training period, but were able to overcome this tendency more quickly to respond appropriately to the object's unity. The only difference between events 1 and 2 is the presence of texture in event 1, which seems to have made perception of object unity more difficult, a counterintuitive result that is at odds with human performance (Johnson & Aslin, 1996). Why would the texture cue have a negative impact on perception of unity in this model?

This result can be explained by considering the training events presented to Model 2, and the nature of learning by connectionist models. Connectionist networks are powerful statistical learners, and can extract even very slight patterns from background noise. This is especially important in the present situation because of the simplicity of the training environment. These networks are 'looking' for subtle cues when, in fact, there is nothing very subtle going on. Because there are an equal number of texture events and non-texture events in the training set, it would seem that accretion and deletion of texture is not a useful cue and that the networks should learn to ignore it. Even though there are an equal number of texture and non-texture events, however, there are slightly more image frames in which the texture module will respond 'yes' (to accretion and deletion of texture) than image frames in which it will respond 'no'. (Recall that an event is made up of a sequence of frames.)

The texture module outputs 1 when the number of texture elements in the current time step is different than it was in the previous time step, and 0 if there is no change. There is exactly the same number of frames with texture as those without texture, but an imbalance arises when we pass from a training event with texture to one without texture. Even though this latter event does not have any texture elements, the texture module will respond initially by signaling that there has been texture deletion in the first time frame, because the network has gone from 'seeing' texture to 'not seeing' texture when passing between events. In other words, the first frame of any event without texture that follows an event with texture will be marked as having texture deletion. This happens in 25% of the events. Each event consists of 14 frames, so the networks experience a texture deletion output (1/14) × 0.25 = 0.02 times more often. That is, approximately 51% of frames involve the texture deletion feature active whereas 49% involve the feature inactive. As noted previously, the enriched training set has many more disjoint events than unified events which implies that there is a slightly higher correlation between the presence of texture deletion and the presence of disjoint objects. For these networks, therefore, the presence of texture deletion is a weak predictor of two objects being present. The correlations that underlie this association are very small, and it is thus a very weak link that only comes into play when the other cues are well balanced. Notably, the networks overcame the tendency to associate texture with disjoint objects eventually in perceiving unity in event 1.

Consider next performance on events to which the networks were not exposed during training. By the end of 8,000 epochs, the networks achieved accurate performance on 14 of these 18 events (77.8%, a higher success rate than the 68.2% success rate of Model 1), suggesting that the additional training events led to greater generalization of knowledge relative to Model 1. Inspection of Figure 8 reveals that the enhanced knowledge of Model 2 incorporates a role for both motion and alignment in perception of object unity. In events 5 and 6, for example, in which there is no motion, unity is perceived accurately by 1,500 epochs. This percept is achieved more quickly than in comparable displays with motion (events 1 and 2), suggesting that motion is a cue that biases *against* unity perception. As in the case of the texture cue described previously, this counterintuitive result (relative to human performance) can be accounted for by appealing to the nature of the training set. Recall that the majority of training events consisted of disjoint objects, and these all contained co-motion as a cue (but not common motion). Motion, therefore, *in the form of co-motion*, became associated with disjoint objects; later in training, common motion (available as a cue in events 1 and 2 in the training set) became associated with unity.

If perception of unity in events 1 and 2 was not achieved primarily on the basis of motion, what cue or cues led to accurate performance? Note that alignment (the combination of parallelism and relatability) was present in these two events but none of the other training events, leading the networks to associate alignment to unity, rather than to disjoint objects. In the absence of motion, therefore, the networks more quickly perceived unity when alignment was available (events 5 and 6). The networks also seemed to use parallelism and relatability separately as cues for unity, even though this led to inaccurate performance: unity was perceived in events 7 and 8, each with relatability but not parallelism, and events 23 and 24, each with parallelism but not relatability. This response pattern was due to the association of each cue to unity in training events 1 and 2. This tendency to perceive unity from parallelism and relatability was overcome, however, with the additional information for disjoint objects provided by the lack of T-junctions in comparable displays (events 9 and 10, and events 25 and 26, respectively). Lack of T-junctions was associated consistently with separate objects during training in events 17, 18, 21 and 22.

In summary, the networks in Model 2 learned to perceive either unified or disjoint objects in a wide range of new events. Performance was superior relative to Model 1, due to the provision of a richer training set. The idiosyncrasies of Model 1 did not characterize performance in Model 2, whose responses were more readily interpreted in light of training experience. The few instances of inaccurate performance in Model 2 were explained by appealing to the nature of learning in connectionist networks, and the limitations of the training environment. These powerful statistical learners extracted regularities that were unique to their training environment and that do not reflect regularities characteristic of the human environment. Increasing the richness and complexity of this environment (thereby bringing it more in line with the infant's environment) should eradicate these spurious correlations. For these networks, therefore, their responses were *not* inaccurate (strictly speaking), given the perceptual environment they were provided.

Using perceptual cues across development

In this section, we explore how associations between cues and percepts build up over learning. In this way we can examine more closely a key concept in investigations of perceptual development: the manner in which an immature perceptual system begins to utilize the sensory information gleaned from the environment to form higher-order representations, and how different cues contribute to these representations at different points in development.

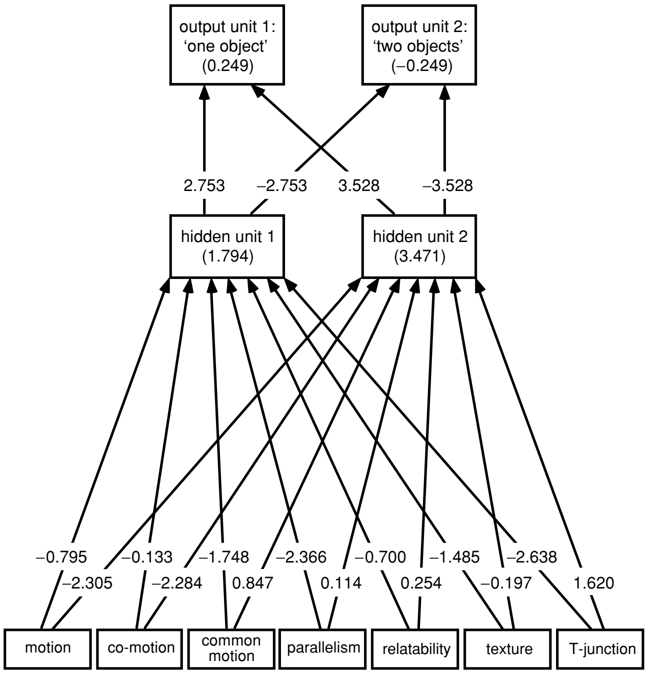


Figure 9 *Connection weights (and resting activations in parentheses), after training, in one of the Model 2 networks. See text for details.*

The simulations reported in the previous section (Model 2) were repeated with the same architecture and training environment, with one exception: an initial investigation of the role of the hidden units in the models described previously showed that only two of the three hidden units played any functional role (i.e. the third hidden unit was redundant). To simplify the analyses of weights in the network, the simulations reported in this section were run using two instead of three hidden units.

Weight matrices after training

A better understanding of how the different cues contribute to the percept of unity can be obtained by looking at the strength of the connection weights in the networks. Figure 9 depicts a schematic view of the connection weights in one network after 8,000 epochs of training with events 1, 2 and 17 to 22. At the top of the figure are the two output units: one of these signals unity, and the other signals disjoint objects (recall that an output across these units of (+1, −1) is scored as a 'one object' or unity response, and an output of (−1, +1) is scored as a 'two objects' response). At the center of the figure are the two hidden units, and at the bottom are the seven perceptual modules. Each hidden unit and each output unit has a variable threshold, which is implemented through the adaptation of its resting level of activation, shown in parentheses. Arrows between boxes depict connection strengths between units. To take account of resting activations, the values shown in Figure 9 consist of the actual connection weight plus the value of the resting activation of the receiving unit. These numbers, therefore, correspond to the extent to which the contribution from the sending unit (i.e. an input unit or a hidden unit) affects the receiving unit (i.e. a hidden unit or an output unit) over and above that receiving unit's resting activation levels.

Inspection of Figure 9 reveals that the weights between the hidden units and output response units are similar: When firing positively, both hidden units will *activate* output unit 1, signaling 'one object', and *deactivate* output unit 2, suppressing the signaling of 'two objects', when firing positively. This is because both hidden units have positive weights to output unit 1 and negative weights to output unit 2. When both hidden units fire negatively, the opposite occurs. When both units agree, therefore, their activation will result in the same output. If the hidden units do not agree, a conflict resolution mechanism has developed: in this particular network, the connection weights from hidden unit 2 are larger than those from hidden unit 1. In case of conflict, then, the response from hidden unit 2 will drive the output response over and above that of hidden unit 1.

Note that output unit 1 has a small positive bias in resting activation (0.249), whereas output unit 2 has a small negative bias in resting activation (−0.249). This tends to favor the network to signal 'one object' (i.e. a (+1, −1) response). This tendency is reinforced by the resting activations of the two hidden units, which are strongly positive in both. In other words, the default for this network, after training, is to perceive unity in the absence of any perceptual evidence.

Inspection of the connection weights between the hidden units and the perceptual modules reveals how the presence or absence of different perceptual cues causes the network's output response. First, note that both hidden units have negative associations with the presence of motion. That is, the presence of motion will tend to cause the hidden units to send negative activation to the output units (especially that of hidden unit 2) and hence will move the network toward a 'two objects' response. As discussed in the previous section, all the training events involve motion, and 6 of the 8 events involve two objects. Motion, therefore, becomes associated with the presence of two objects. A more subtle point follows from this. Because motion is present in all the training events, negative activation is always flowing into the hidden units during learning. When paired with the positive bias that is always present, the motion cue effectively resets the baseline activation to below zero. So, although there is an initial bias toward producing a 'one object' response, that bias is reversed in the presence of motion. As discussed subsequently, the presence of other cues negatively and positively associated with either hidden unit can tip that unit's response, and the network's response, in one direction or the other.

Next, note that the other cues will each tend to have different effects on hidden unit 2, the stronger of the two hidden units. Recall that this hidden unit is initially biased to produce a 'one object' response, and that an input of strong negative activation will produce the opposite response. Hidden unit 2 has a strong negative weight with the co-motion cue (–2.284). Along with the motion cue, therefore, the presence of co-motion tends to produce a 'two objects' response. T-junctions are most strongly associated with the presence of one object (1.620), followed by common motion (0.847), relatability (0.254) and parallelism (0.114). Finally, for this network, the texture is a weak predictor of two objects being present. Note that hidden unit 2 has picked out many of the cues that we would normally see as predicting unity, and that both relatability and parallelism provide independent contributions to the 'one object' response.

Hidden unit 1, the weaker of the two hidden units, has developed a completely different set of associations. Specifically, this unit has developed negative associations with *all* cues. As a result, as long as something is present on a display, its response will counteract the initial bias towards responding with unity by attempting to cause a 'two objects' response. This embodies the fact that the large majority of training events involve two objects. Hidden unit 2 must thus fire strongly to bring the network to a 'one object' response. Any weak response from hidden unit 2 (e.g. when cues provide conflicting information) allows hidden unit 1 to activate the output unit 2 by firing negatively.

Examination of these connection weights, therefore, illustrates how the network is able to combine evidence in a complex fashion from a series of perceptual modules to make a unity prediction. In particular, no single cue is sufficient to perceive unity, and cues take on different degrees of importance in different contexts. In the next section we explore how these associations build up with learning.

Learning cue associations

Figure 10 shows the connection weights between the outputs of the perceptual modules and the hidden units across development. Connection weights are represented by squares. The larger the magnitude of the weight, the

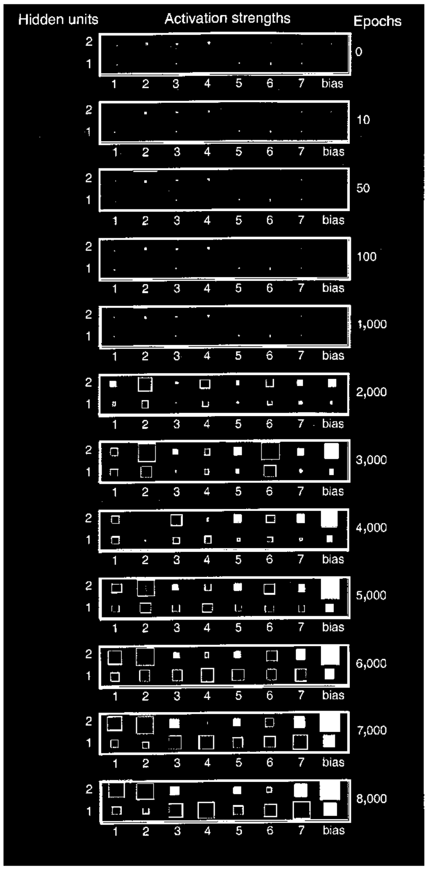


Figure 10 *Development of connection weights between the seven perceptual modules and bias unit, and the two hidden units. Right to left: 1. Motion. 2. Co-motion. 3. Common motion. 4. Parallelism. 5. Relatability. 6. Texture. 7. T-junctions. Top to bottom: Activation strengths of each connection after varying numbers of epochs.*

larger the square; positive values are in white and negative values are in black. Each of the 12 panels depicts the connection matrix at different points in development (top to bottom: epochs 0, 10, 50, 100, 1,000, 2,000, 3,000, 4,000, 5,000, 6,000, 7,000 and 8,000).

For epochs 0 to 1,000 the weights from the perceptual modules remain close to zero. Initially, then, the network does not utilize information from *any* of the perceptual modules in producing its response. That response (which is predominantly a 'two object' response) is triggered by a strong early bias which is set by adjusting the output units' resting activation levels. Because most of the events in the learning environment arise from two objects, a quick solution to predicting when there is one or two objects is to always predict two. This is correct in many cases and does not require the assimilation of multiple perceptual cues. Of course, this is not optimal performance, and the network uses this strategy only as an initial solution.

At 2,000 epochs, the network begins to rely on different perceptual cues to produce its response and the basic structure of the mature weight matrix (discussed in the previous section) begins to appear. Hidden unit 2 acquires a fairly large positive bias (thereby biasing it to produce a unity response). Motion, common motion, relatability and T-junctions have become positively associated with this unit, and hence these all become cues that signal a single unified object. In contrast, co-motion and texture become strongly associated with the percept of two objects. At this point, hidden unit 1 has the same (though weaker) pattern of associations between perceptual cues and the output response as hidden unit 2. Initially, then, hidden unit 1 is redundant.

By 3,000 epochs, the weights have continued to grow in magnitude, but have remained in essentially the same pattern as at 2,000 epochs. A weaker response from the perceptual modules, therefore, will trigger the same responses from the network, which is becoming more sensitive to weaker forms of the same evidence in interpreting an ambiguous stimulus. By 4,000 epochs, hidden unit 1 has emerged with it own role: all the connections between the perceptual modules and this unit have become negative. At this point, hidden unit 1 acts to bias the network against a unity response, thereby ensuring that hidden unit 2 elicits a unity response only when there is strong evidence to do so.

From 5,000 epochs onwards, the pattern of associations between the perceptual modules and the hidden units remains stable. The changes with development involve an increase in the magnitude of the weights, leading to an increased sensitivity to the relevant cues. By epoch 6,000, parallelism decreases in association with hidden unit 2 until it reaches a negligible level at 8,000 epochs.[3] The texture cue also weakens its association with hidden unit 2. In contrast, all weights linking the perceptual modules with hidden unit 1 continue to increase in magnitude.

In summary, throughout development the network progressively relies on an increasing range of cues to determine its response. The connection weights embody the association between perceptual cues and unity in both ambiguous and unambiguous contexts. The challenge for the network is to discover a set of weights that

[3] Interestingly, parallelism did not become associated with unity during any time of this network's development, unlike the majority of the Model 2 networks reported previously. This is the only instance in which the present network and the previous networks differed with respect to which cues predicted unity. The difference may lie in the somewhat stochastic nature of the development of connectionist networks, due to variations in starting conditions, randomized training schedules, and the fact that there is no unique set of cue combinations that can be used to perceive unity correctly.

is consistent with both contexts, and is consistent with both one object and two object events. Many of the associations that emerge reflect the fact that the majority of events this network experiences are unambiguous (fully visible) and arise from two distinct moving objects. These sorts of events are the first that are learned in the network's mediated route. Only later in development (4,000 epochs and afterward) does the mediated route begin to learn associations that allow it to function with ambiguous events.

An important finding from the analyses in this section and the previous section is that no single cue is necessary or sufficient to the perception of unity. Instead, the importance of a cue depends on the context in which it appears. Individual cues do not maintain constant importance as markers of unity across all events, but acquire more or less importance in accord with other available information.

General discussion

Our goal in designing connectionist models of the development of perception of object unity during partial occlusion was to explore whether veridical percepts (i.e. matching adult human performance) could arise in the absence of prior knowledge of occlusion, and instead emerge as a result of pre-established lower-level perceptual skills, learning algorithms and specific kinds of experience. The answer to this question is clearly affirmative. Even Model 1, which was trained in a relatively impoverished perceptual environment (i.e. with a small sample of the events), achieved good performance with novel events, in perceiving unity when it was potentially ambiguous. Model 2 experienced a richer perceptual environment during training and achieved even better performance.

The account of object perception that ensues from these findings is consistent with a general law of parsimony: a system of higher-order knowledge develops over time from a series of lower-level abilities, in accord with the structure of its input. Why is parsimony to be desired? We argue that parsimony is a central pillar of science: the best explanations are the simplest, and those that concur with known quantities (e.g. observable behaviors). This view, of course, is not new (e.g. Morgan, 1903). As discussed previously, there is consistent evidence that neonates have clear visual preferences (Slater, 1995). They direct attention in particular to salient edges and motion, and thereby gain experience with two kinds of visual information that are especially important in parsing the optic array (Johnson, 1997). This is exactly the right learning environment from which an early ability to segment figure from ground could progress to an ability to perceive partial occlusion. From there, more sophisticated percepts emerge, such as the ability to represent objects while fully occluded, but these percepts too are fragile initially (Johnson, Bremner, Slater, Mason, Foster & Cheshire 2002; Johnson, 2000, 2001). Such an account does not rely on the positing of innate knowledge in infants to explain the development of object percepts. Our models are consistent with this account. They implement an initial *sensitivity* to visual information (that has been shown to support veridical responses to partly occluded objects). They then show how these visual cues can be combined to produce a representation of the ambiguous state of a partly occluded object: a representation that was not present in any form prior to experience with the environment.

It might be argued that the network has not actually acquired any new representation of unity because there was a set of 'unity' nodes present from the onset (i.e. the output units that we labeled externally as corresponding to unity). It is true that this is a network with a fixed architecture, and therefore a predetermined number of representational nodes (see Mareschal & Shultz, 1996 for a discussion of networks with a variable number of nodes), but we believe that the network has nevertheless acquired a new level of representation. The output is initially not linked to any perceptual input. At this early stage, although we are labeling the output as 'unity', it has no *semantic* content because it is not grounded in any perceptual experience and consequentially cannot be used to refer to any event in the environment. This is followed by a stage in which the output is linked to certain perceptual cues that lead the network to respond to the ambiguous stimulus in Figure 1a as though it were caused by two broken but partly occluded rods. At this stage the nodes have acquired semantic content that is grounded in perceptual input and leads them to classify ambiguous events as 'not unified'. In other words, the network now has a representation of what is unified and what is not unified, although this representation differs from the normal adult representation. Finally, at the end of training, the unity output nodes have become associated to yet another set of perceptual cues. The response is still perceptually grounded (and therefore can be used to refer to something in the environment). However, the semantic content of the response has changed. Now, the network's unity response when presented with an ambiguous stimulus is the same as that of adults, and the ambiguous stimulus is perceived as corresponding to a single unbroken but partly occluded rod. Within the constraints imposed by the limited perceptual cues available, the network's representation of 'unity' has acquired the same semantic content as that of adults.

The importance of parsimony in theory building is also underscored by considering that the design of the visual system is matched to the statistics of the visual environment. That is, the information processing strategies employed by the visual system exploit structure inherent to natural visual scenes (Field, 1999a, 1999b). These strategies allow for vision (and, presumably, other sensory systems) to capitalize on redundancies in the stimulus input to focus on areas of contrast and change (cf. Marr, 1982). It seems likely, therefore, that visual development will also concur with input from the visual environment.

How well does model performance match human performance?

The models were designed to produce one of three outputs: perception of object unity, perception of disjoint objects, or an indeterminate response. The networks' initial responses indicated perception of disjoint objects in all portions of all events. This strategy was not built into the models, but was acquired very quickly via the available learning mechanisms. Performance improved with the emergence of unity percepts on the ambiguous portions of those events in which unity was present, but not directly perceivable. Over the course of learning, the networks determined which cues and which cue combinations were most reliably associated with unity.

The event in Figure 1a is truly ambiguous in that there is no necessary or sufficient reason to infer that it arises from an occluder and either a single, partially occluded rod or two partially occluded rods that happen to be aligned and moving together. An unbiased rational agent would be unable to decide whether the rod was broken or not. This is not the case for humans, however. As noted previously, neonates perceive this stimulus as arising from two rods, whereas 4-month-olds and adults tend to perceive it as arising from a single rod. These responses imply that both newborns and 4-month-olds are endowed with 'inductive biases', or information processing constraints that bias a system to select preferentially one response among a set of equivalent outputs (Mitchell, 1997). Note that neither response is 'correct' in the sense of being a uniquely valid inference from the information available in the stimulus, because both are equally compatible with the available data. An important implication of our models is that these inductive biases do not have to be built in. Inductive biases can arise – and change – as a result of experience with a particular environment.

The incipient tendency to perceive disjoint objects, the gradual development of unity percepts and the discovery of multiple cues to unity are strikingly similar to human performance. Initially, the networks used their existing perceptual skills to segregate the events into constituent object surfaces based on what was directly perceived, although these segregation skills were not pre-specified. After repeated exposure to events in which objects were seen first fully visible and then underwent partial occlusion, percepts of unity emerged that matched human adult percepts. The use of cues to perceive unity changed with exposure. Both these latter patterns of performance are also observed in infants. Overall, then, the progression from responding to partly occluded objects as disjoint objects to perceiving object unity characterizes both our connectionist models and human infants, implying common developmental mechanisms. This lends plausibility to our account of these mechanisms as arising from an early perceptual competence (but see footnote 2 again) and experience viewing objects as they become progressively occluded and disoccluded.

The models employed simplistic perceptual modules and experienced a relatively impoverished environment (as compared to the natural world). Thus, it is unlikely that specific predictions about infants' reliance on particular cues can be derived from these models. This is because the infants experience a much richer environment than the networks did. Nevertheless, the models embodied the computational principles by which human infants might learn. They successfully show how associative mechanisms can be used to combine perceptual cues in such a way as to derive a unity response similar to that of adults, from a perceptually ambiguous event.

It is also notable that early in training, the models made no use of perceptual information in making a response. After 2,000 epochs, however, there was a relatively sudden emergence of cue use whose basic pattern remained largely unchanged throughout development. That is, there was a sudden shift in performance, strongly resembling a stage (cf. Elman *et al.*, 1996).

Of course, there are also important differences in how these connectionist models and humans use information to parse scenes containing partly occluded objects. Our models were not intended to tell us about which cues infants use to perceive unity (indeed our selection of input cues was driven by previous experimental studies with infants). They were designed to test the prediction that perceptual sensitivity and association lead to a response bias towards ambiguous stimuli that has been interpreted as evidence of object knowledge. Texture, for example, was not used by the models as a depth cue, as is the case for human adults and infants (Gibson, 1979; Johnson & Aslin, 1996, 2000).

What do we mean by knowledge?

Computational models force the user to be explicit about what is meant by the term 'knowledge'. In the present context, knowledge of object unity is inferred from the models' and from the infants' behavior. In the models, this knowledge is embodied in the connection weights between units in the networks, and no propositional or other explicit representation of the task is required. The weights internalize the time-averaged co-occurrence of different features in the environment. Our models suggest that this representational format seems sufficient to account for performance on the partly occluded object tasks depicted in Figure 1. Moreover, what has been described as a process of 'perceptual inference' to account for the behavior of 4-month-olds is explicitly instantiated here, through the propagation of activation in a neural network and not through the syntactic manipulation of propositions.

What is innate and what is learned?

As noted previously, our models were not intended to explain visual development, nor should they be taken as a general model of how all knowledge is acquired. Rather, they are best viewed as an existence proof: object knowledge can arise from an interaction of lower-level perceptual skills and learning, if given the proper experience in a structured environment.

Is our model purely empiricist? The answer is a clear no, because the starting point for development was specified. Neither, however, is it purely nativist, because something completely novel arose during training: object knowledge apart from the original perceptual sensitivities. Ultimately, any innate vs learned dichotomy fails to provide explanatory power in addressing developmental issues. We argued previously that our models are consistent with a general law of parsimony, because they are consistent with known facts: infants are born with preferences for important visual cues, they learn and they live in a structured visual world. But this approach does not view the developing human as a 'black box' or *tabula rasa* in which internal mechanisms are unobservable, and therefore unimportant. On the contrary: understanding the internal structure of our models, during and after development, is crucial for the interpretation of their performance.

Grounding our discussion in an implemented computational model makes explicit the requirements for a learning account of perception of object unity. We have shown that, given a set of seven perceptual cues, a simple associative learning system can acquire a bias to perceive Figure 1a as a single partly occluded object. To push a learning argument further, one would need to identify under what set of constraints these perceptual modules would emerge. In this way, real progress can be made in identifying what information needs to be available to the system for learning to get off the ground.

The models in this paper demonstrate how the perception of unity could be mediated by available information in the absence of direct evidence. We do not wish to claim that there is anything special about unity in this case: it is relatively straightforward to generalize this account to other perceptual cues. Any one of these cues could be mediated by indirect associations with other cues. A more complex network could be devised, in which, if one cue could not be computed, its association with other computable cues could be used to derive a value for that cue. However, whether the resulting network would be computationally tractable is an open question.

Appendix: Details of the model architecture

The networks were designed with an input layer (the perceptual modules), an output layer (the response units) and a layer of hidden units between the perceptual modules and the response units. They were exposed to a simple perceptual environment. There were no direct connections from the perceptual modules to the output; all cue relations, therefore, had to be encoded across the hidden units. Networks began with random initial weights (mean = 0.0, range = 0.01). During training, the weights between the perceptual modules, hidden units and the output units were updated at every epoch (i.e. image presentation) using the backpropagation algorithm[4] (Chauvin & Rumelhart, 1995) with a learning rate of 0.5, momentum of 0.03. The memory parameter (μ) was set to 0.4. A network's performance was tested by presenting it with events consisting of the ambiguous segment of the trajectory only. In other words, during testing, the networks could not use information available in the unambiguous segment of the trajectory to derive unity.

Output representations

The unity response was coded across two linear output units with activations ranging across the interval (–1, +1). An output activation pair of (+1, –1) signified that the surfaces were unified, and (–1, +1) signified that the surfaces were not unified. A response of (+1, +1) or (–1, –1) was interpreted as indeterminate.

[4] We have no necessary commitment to backpropagation. Any connectionist algorithm that implements gradient descent search in multi-layered networks could be used equally well, such as the leabra algorithm (O'Reilly, 1996, 1998).

Input representations

The input consisted of a 196-bit vector mapping all the units on a 14 × 14 grid. In the center of the grid was a 4 × 4-unit occluder. All units corresponding to the position of the occluder and visible object parts were given a value of 1. When background texture was present, all other units on the display were given a value of 0 or 0.2. Units with values of 0.2 corresponded to positions on which there was a texture element (i.e. the dots seen in Figure 4). Each event consisted of a sequence of 14 snapshots in which the object moved progressively across the display.

The perceptual modules

The motion detection module

This module compared the current image to the previous image (Figure 11a). If there was a difference between the images, the presence of motion was computed. Input during a particular time step (e.g. 'time1') was copied to a memory buffer (Prev.Input), and input from the next time step ('time2') was copied to a second module (Curr.Input). A layer of hidden units (Diff.Input) within the module computed the step by step difference between Prev.Input and Curr.Input. The output unit then summed the activity across the hidden layer, and output 1 if there were any non-zero values or 0 if there were no non-zero values.

The co-motion module

This module split the display into two halves and computed whether there was motion simultaneously in the upper half and the lower half. To accomplish this, two motion detection modules (described previously) were employed, one devoted to each half, and their outputs fed to a third module that determines whether both are active concurrently. If both were active, the output unit output 1, or 0 otherwise.

The common motion module

This module split the display into two halves and computed whether there was the same motion in the upper half and the lower half (Figure 11b). A Diff.Input vector was generated for both the top and bottom halves of the display (as in the motion detection module described previously). Recall that the contents of Diff.Input were the current input minus the input at the previous time step. Features that had not moved cancelled out and left a value of 0 at their associated positions in Diff.Input.

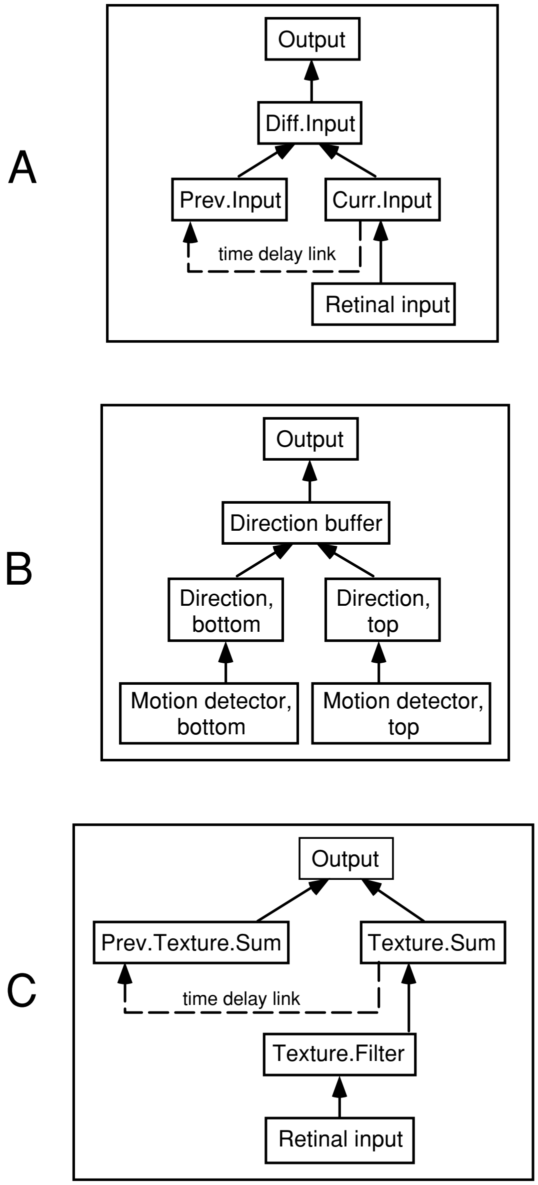


Figure 11 *Architecture of three perceptual modules. A. Motion module. B. Co-motion module. C. Texture module. See text for details.*

Positions that were newly occupied with the current input were provided with positive activation and positions that were occupied at the previous time step, but were no longer occupied, were provided with negative activation. The direction of motion was determined by observing the relative position of positive and negative activation. A Direction Buffer computed a weighted sum of the negative and positive activation (in which locations along the horizontal retinal axis were increasingly weighted from left to right) for each of the top and bottom halves of the display. A positive sum indicated that the object was moving to the right, a negative sum indicated that the object was moving to the left, and 0 indicated that there was no motion. Comparing these two values for the top and bottom halves of the display allowed the module to compute whether there is the same kind of motion in the top and bottom halves of the display. If there was common motion the module output 1, or 0 otherwise.

The parallelism module

This module computed an approximation to the tangent of the angle that the object's axis of principle length made with the horizontal, for both the upper and lower halves of the display, and compared these two values. A Cartesian coordinate system was set up by weighting the columns and the rows of the display according to the position of the row and column with respect to an origin at the center of the display. The X- and Y-components of object segments in the upper and lower halves of the display were computed within this coordinate system. The ratio of the Y-component over the X-component was used as an approximation to the tangent of the angle. If the tangents in the top and bottom halves were equal, the module output 1, or 0 otherwise.

The relatability module

This module computed whether the extension of the axis of principle length for objects in the upper and lower halves of the display would intersect. The display was split into two halves. The bottom half was copied on to the top half and the occluder values subtracted. If axes converged while moving up the display, then the objects were relatable and the module output 1, or 0 otherwise.

The texture module

This module summed the number of texture dots in the input image and compared it to the number of dots in the previous image (Figure 11c). If there was a difference in the number of dots in the two images, then there had been accretion or deletion of texture elements. The input was passed via one-to-one connections to a layer of hidden units (Texture.Filter). A unit in this layer was activated if its corresponding position on the display contained a texture element. The output of Texture. Filter was then passed on to a single unit (Texture.Sum) that summed the activation across the whole layer. Because each unit in Texture.Filter had an activation of 1 if there was a texture element, the sum of all the units was equivalent to the number of texture elements present at that time step on the display. This value was passed into a memory buffer (Past.Texture.Sum) for use at the next time step. The values of Past.Texture.Sum and Texture.Sum were passed to an output unit that computed the difference between the two. The output unit responded with a 1 if the difference was not 0 (i.e. there had been texture accretion or deletion), or a 0 if there was no difference (i.e. no background texture was present).

The T-junction module

This module focused on the area immediately above and below the edge of the occluding screen and computed whether there was a gap along these edges. If there was a gap, the absence of T-junctions was computed and the module output a 0, or a 1 if there was no gap.

The cues detected by the perceptual modules were primitives, and other cues were computed as combinations of these primitives. For example, collinearity was indicated by positive responses from both the parallelism and relatability modules. Parallelism and relatability are more primitive cues than collinearity because the latter cannot be computed without the former, whereas the converse is not true (i.e. both parallelism and relatability can be computed independently of collinearity). Also, co-motion, common motion, parallelism and relatability can only be computed when there is more than one possible object present (i.e. when there are two objects in either an ambiguous or unambiguous situation).

Acknowledgements

This research was supported by NSF grant BCS-9910779, ESRC grant R000238340, ESRC grant R000239112 and European Commission HP-RTN grant CT-2000-00065. Portions of this research were presented at the 1998 meeting of the International Society for Infant Studies, Atlanta, GA and the 1999 meeting of the Cognitive Science Society, Vancouver, BC.

References

Baldi, P., Chauvin, Y., & Hornik, K. (1995). Backpropagation and unsupervised learning in linear networks. In Y. Chauvin & D.E. Rumelhart (Eds.), *Backpropagation: Theory, architectures, and applications* (pp. 389–432). Hillsdale, NJ: Erlbaum.

Banton, T., & Bertenthal, B.I. (1997). Multiple developmental pathways for motion processing. *Optometry and Vision Science*, **74**, 751–760.

Bornstein, M.H. (1985). Habituation of attention as a measure of visual information processing in human infants: summary, systematization, and synthesis. In G. Gottlieb &

N.A. Krasnegor (Eds.), *Measurement of audition and vision in the first year of postnatal life: A methodological overview* (pp. 253–300). Norwood, NJ: Ablex.

Chauvin, Y., & Rumelhart, D.E. (1995). *Backpropagation: Theory, architectures, and applications*. Hillsdale, NJ: Erlbaum.

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.

Field, D.J. (1999a). What is the goal of sensory coding? *Neural Computation*, **6**, 559–601.

Field, D.J. (1999b). Wavelets, vision and the statistics of natural scenes. *Philosophical Transactions of the Royal Society of London A*, **357**, 2527–2542.

Gibson, J.J. (1979). *The ecological approach to visual perception*. Hillsdale, NJ: Erlbaum.

James, W. (1890). *The principles of psychology*. New York: Holt.

Johnson, S.P. (1997). Young infants' perception of object unity: implications for development of attentional and cognitive skills. *Current Directions in Psychological Science*, **6**, 5–11.

Johnson, S.P. (2000). The development of visual surface perception: insights into the ontogeny of knowledge. In C. Rovee-Collier, L. Lipsitt & H. Hayne (Eds.), *Progress in infancy research* (Vol. 1, pp. 113–154). Mahwah, NJ: Erlbaum.

Johnson, S.P. (2001). Visual development in human infants: binding features, surfaces, and objects. *Visual Cognition*, **8**, 565–578.

Johnson, S.P., & Aslin, R.N. (1995). Perception of object unity in 2-month-old infants. *Developmental Psychology*, **31**, 739–745.

Johnson, S.P., & Aslin, R.N. (1996). Perception of object unity in young infants: the roles of motion, depth, and orientation. *Cognitive Development*, **11**, 161–180.

Johnson, S.P., & Aslin, R.N. (2000). Infants' perception of transparency. *Developmental Psychology*, **36**, 808–816.

Johnson, S.P., Bremner, J.G., Slater, A., & Mason, U. (2000). The role of good form in young infants' Perception of partly occluded objects. *Journal of Experimental Child Psychology*, **76**, 1–25.

Johnson, S.P., Bremner, J.G., Slater, A.M., Mason, U., Foster, K., & Cheshire, A. (2002). Infants' perception of object trajectories. Manuscript submitted for publication.

Johnson, S.P., & Náñez, J.E. (1995). Young infants' perception of object unity in two-dimensional displays. *Infant Behavior and Development*, **18**, 133–143.

Jusczyk, P.W., Johnson, S.P., Spelke, E.S., & Kennedy, L.J. (1999). Synchronous change and perception of object unity: evidence from adults and infants. *Cognition*, **71**, 257–288.

Kellman, P.J. (1996). The origins of object perception. In E. Carterette & M. Friedman (Series Eds.) & R. Gelman & T. Au (Vol. Eds.), *Handbook of perception and cognition: Perceptual and cognitive development* (2nd edn., pp. 3–48). San Diego: Academic Press.

Kellman, P.J., & Shipley, T.F. (1991). A theory of visual interpolation in object perception. *Cognitive Psychology*, **23**, 141–221.

Kellman, P.J., & Spelke, E.S. (1983). Perception of partly occluded objects in infancy. *Cognitive Psychology*, **15**, 483–524.

Kellman, P.J., Spelke, E.S., & Short, K.R. (1987). Infant perception of object unity from translatory motion in depth and vertical translation. *Child Development*, **57**, 72–86.

LaPlante, D.P., Orr, R.R., Neville, K., Vorkapich, L., & Sasso, D. (1996). Discrimination of stimulus rotation by newborns. *Infant Behavior and Development*, **19**, 271–279.

Laplante, D.P., Orr, R.R., Vorkapich, L., & Neville, K.E. (2000). Multiple dimension processing by newborns. *International Journal of Behavioral Development*, **24**, 231–240.

Mareschal, D. (2001). Connectionist methods in infancy research. In J. Fagen & H. Hayne (Eds.), *Progress in infancy research* (Vol. 2, pp. 71–119). Mahwah, NJ: Erlbaum.

Mareschal, D., French, R.M., & Quinn, P.C. (2000). A connectionist account of asymmetric category learning in early infancy. *Developmental Psychology*, **36**, 635–645.

Mareschal, D., & Shultz, T.R. (1996). Generative connectionist networks and constructivist cognitive development. *Cognitive Development*, **11**, 571–605.

Marr, D. (1982). *Vision*. San Francisco: Freeman.

Mitchell, T.M. (1997). *Machine learning*. New York: McGraw-Hill.

Morgan, C.L. (1903). *Introduction to comparative psychology*. New York: Scribner.

Munakata, Y., McClelland, J.L., Johnson, M.H., & 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.

Nakissa, R.C., & Plunkett, K. (1998). Evolution of a rapidly learned representation for speech. *Language and Cognitive Processes*, **13**, 105–127.

Náñez, J.S. (1988). Perception of impending collision in 3- to 6-week-old infants. *Infant Behavior and Development*, **11**, 447–463.

Needham, A., Baillargeon, R., & Kaufman, L. (1997). Object segregation in infancy. In C. Rovee-Collier & L. Lipsitt (Eds.), *Advances in infancy research* (Vol. 11, pp. 1–44). Norwood, NJ: Ablex.

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

O'Reilly, R.C. (1998). Six principles for biologically-based computational models of cortical cognition. *Trends in Cognitive Sciences*, **2**, 455–462.

Piaget, J. (1952). *The origins of intelligence in children*. New York: International Universities Press.

Piaget, J. (1954). *The construction of reality in the child*. New York: Basic Books.

Slater, A. (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.

Slater, A., Bremner, G., Johnson, S.P., Sherwood, P., Hayes, R., & Brown, E. (2000). Newborn infants' preference for attractive faces: the role of internal and external facial features. *Infancy*, **1**, 265–274.

Slater, A., Brown, E., & Badenoch, M. (1997). Intermodal perception at birth: newborn infants' memory for arbitrary auditory-visual pairings. *Early Development and Parenting*, **6**, 99–104.

Slater, A., Johnson, S.P., Brown, E., & Badenoch, M. (1996). Newborn infants' perception of partly occluded objects. *Infant Behavior and Development*, **19**, 145–148.

Slater, A., Mattock, A., & Brown, E. (1990). Size constancy at birth: newborn infants' responses to retinal and real size. *Journal of Experimental Child Psychology*, **49**, 314–322.

Slater, A., Morison, V., Somers, M., Mattock, A., Brown, E., & Taylor, D. (1990). Newborn and older infants' perception of partly occluded objects. *Infant Behavior and Development*, **13**, 33–49.

Slater, A., Von der Schulenburg, C., Brown, E., Badenoch, M., Butterworth, G., Parsons, S., & Samuels, C. (1998). Newborn infants prefer attractive faces. *Infant Behavior and Development*, **21**, 345–354.

Smith, W.C., Johnson, S.P., & Spelke, E.S. (in press). Motion and edge sensitivity in perception of object unity. *Cognitive Psychology*.

Spelke, E.S. (1990). Principles of object perception. *Cognitive Science*, **14**, 29–56.

Spelke, E.S. (1994). Initial knowledge: six suggestions. *Cognition*, **50**, 431–445.

Spelke, E.S., & Van de Walle, G. (1993). Perceiving and reasoning about objects: insights from infants. In N. Eilan, R.A. McCarthy & B. Brewer (Eds.), *Spatial representation: Problems in philosophy and psychology* (pp. 132–161). Oxford: Blackwell.

Spillman, L., & Werner, J.S. (1990). *Visual perception: The neurophysiological foundations*. San Diego: Academic Press.

Walton, G.E., & Bower, T.G.R. (1993). Newborns form 'prototypes' in less than 1 minute. *Psychological Science*, **4**, 203–205.

Walton, G.E., Bower, N.J., & Bower, T.G.R. (1992). Recognition of familiar faces by newborns. *Infant Behavior and Development*, **15**, 265–269.

Wattam-Bell, J. (1991). Development of motion-specific cortical responses in infancy. *Vision Research*, **31**, 287–297.

Wattam-Bell, J. (1996a). Visual motion processing in 1-month-old infants: habituation experiments. *Vision Research*, **36**, 1679–1685.

Wattam-Bell, J. (1996b). Visual motion processing in 1-month-old infants: preferential looking experiments. *Vision Research*, **36**, 1671–1677.

Received: 29 September 2000
Accepted: 1 February 2001

COMMENTARIES

Connectionist models of infant perceptual and cognitive development

Leslie B. Cohen[1] **and Harold H. Chaput**[2]

1. Department of Psychology, University of Texas at Austin, USA
2. Department of Computer Science, University of Texas at Austin, USA

In recent years an exciting innovation has occurred in the study of early development. Connectionist models have begun to appear on a variety of aspects of infant perception, cognition and language. These models have included simulations of object permanence (Munakata, McClelland, Johnson & Siegler, 1997); categorization (Mareschal & French, 2000); causal perception (Chaput & Cohen, in press); early word learning (Schafer & Mareschal, 2001); and simple rule learning (Shultz & Bale, in press). Continuing in this tradition, Mareschal and Johnson now present a connectionist model of a cornerstone in infant visual perception, the perception of object unity. This is an ideal topic to model for a number of reasons: as the authors note, the perception of object unity is subject to a continuing theoretical debate about the role of innate core knowledge versus early experience and learning; there is also a wealth of empirical evidence with infants upon which to draw; and some of that empirical evidence suggests a developmental change in how occluded objects (or object parts) are perceived.

Mareschal and Johnson's model includes several attractive features. It is essentially a constructive model that builds the percept of one versus two objects from a set of lower-order perceptual cues. Most, if not all, of these perceptual cues (e.g. the presence of motion, texture or T-junctions) make sense, and they are probably available to young infants. The model learns from environmental experience and generalizes to novel instances; it also produces some, albeit weak, evidence for a developmental change.

Nevertheless, the present model, like many earlier connectionist models, serves mainly as an existence proof, meant to demonstrate that such a system *can* learn some concept. In that sense it provides a logical counterargument to postulations of certain types of innate core knowledge. But the field has now advanced to the point where it is important to go beyond existence proofs and attempt to model more closely how these concepts are actually learned by infants and how developmental changes in these concepts actually occur.

We too are working on a connectionist model of infant cognitive development (Chaput and Cohen, in press), and we would be the first to admit that our own attempts influence our view of other models. Nevertheless, based on this experience we would like to raise a few general issues we have confronted that also apply to the Mareschal and Johnson model. These are (1) choice of a connectionist architecture, (2) modeling habituation vs long-term experience, (3) the realism of the training environment, and (4) modeling developmental change.

A growing trend in connectionist models of infant development is the use of architectures that are unsupervised and self-organizing. We agree with this trend, as it avoids many of the problems associated with the use of an external 'target' against which the model's output is compared. It is unclear where, in infants, a target would come from, or how such a target would be chosen in a consistent way throughout development and across domains. Auto-associative networks answer this question quite nicely by making the target identical to the input. In our own work, we avoid the question entirely by using Self-Organizing Maps (Kohonen, 1997), which make no use of targets and are, thus, unsupervised. Although Mareschal and Johnson claim that their model is unsupervised, their use of an external target, via the 'Direct Route', makes their model supervised by definition. And the source of this target is not actually the input vector (as it would be in an auto-associative network) but the result of an algorithmic manipulation of environmental features not present in the input vector. So their model is not really self-organizing, either. In

Address for correspondence: Leslie B. Cohen, Department of Psychology, Mezes Hall 330, University of Texas, Austin, TX 78712; e-mail: cohen@psy.utexas.edu or chaput@cs.utexas.edu

spite of these possible criticisms, to us, the specific architecture of the model is less important than the results it produces.

Most infant experiments employ a habituation technique to explore perceptual or cognitive development. This includes the research on object unity cited by Mareschal and Johnson; research they presumably are trying to model. But habituation can be used either to assess short-term learning *or* to probe existing knowledge structures that are the result of long-term experience. We believe it is important for the model to be explicit about these two types of learning, as our model is. The distinction is unclear in the Mareschal and Johnson model. If their training is meant to represent short-term habituation, then it is unlike what infants have received in actual experiments since it includes exposure to an unoccluded event in addition to an occluded event. On the other hand if the training is meant to represent long-term experience, then it is unclear how habituation is represented in their model at all.

One can also ask how realistic or artificial that long-term training is. We believe, as Mareschal and Johnson do, that infants are able to learn concepts by processing information they receive from the environment. But we have doubts about the plausibility of the environment provided in Mareschal and Johnson's models, especially Model 2. The authors posit that an infant's bias to perceive a partially occluded rod as two separate rods is the result of an 'inductive bias' which is present, not only in 4-month-olds, but also in neonates. They go on to argue that this 'inductive bias' comes about 'as a result of experience with a particular environment'. Earlier, though, they note that Model 2 prefers the two-rod response 'because most of the events in the learning environment arise from two objects'. So we wonder if they might be 'giving away the answer' by skewing the environment in such a way as to facilitate a certain outcome. Our intuition is that an infant's environment is more likely to contain whole rods than co-moving, parallel, relatable rod segments.

Finally, Mareschal and Johnson's model is only one part of a much larger story of cognitive development, and we would be interested in seeing if their model can cover more of this story. Specifically, with regard to object unity, Eizenman and Bertenthal (1998) performed a similar bar-and-occlusion study with infants; only instead of moving the bar laterally behind the occluder, they *rotated* the bar. Rotation made the perceptual job more difficult for the infant, and they found that 4-month-olds tended to regress to a two-bar bias, whereas 6-month-olds finally regained the single-bar bias. Their results suggest the presence of an additional transition in the development of object unity, and we are curious to see if Mareschal and Johnson's model can capture this transition as well.

We also believe the set of developmental transitions found in infants' perception of object unity represent but a single example of a more general set of principles that appear to be at work across many domains throughout early cognitive development (Cohen, 1998; Cohen & Cashon, 2001). These principles provide a constructivist view of cognitive development amenable to connectionist modeling. Our own work (Chaput & Cohen, in press) is an initial attempt to address this broader picture with a single connectionist architecture that adheres to these principles.

We hope these comments are taken in the positive way they are intended. We believe the authors' approach, along with the other approaches we have mentioned, such as auto-associative networks, cascade correlation models, or our own hierarchically arranged self-organizing maps, hold considerable promise for explaining the development of early perceptual and cognitive ability. We are pleased to see that connectionist models are progressing beyond existence proofs and attempting more complete and potentially compelling explanations. In the end these models will be evaluated by more than their ability to simulate existing empirical data. They will also be required to generate new developmental predictions that are then confirmed empirically. Once they reach that stage they clearly will be making a major contribution to our understanding of early perceptual and cognitive development.

Acknowledgement

Preparation of this paper was supported in part by NIH grant HD-23397 to the first author from the National Institute of Child Health and Human Development.

References

Chaput, H.H., & Cohen, L.B. (in press). A model of infant causal perception and its development. *Proceedings of the 2001 Cognitive Science Society Meeting*, Edinburgh.

Cohen, L.B. (1998). An information-processing approach to infant perception and cognition. In F. Simion & G. Butterworth (Eds.), *The development of sensory, motor, and cognitive capacities in early infancy* (pp. 277–300). Hove, East Sussex: Psychology Press.

Cohen, L.B., & Cashon, C.H. (2001). Infant object segregation implies information integration. *Journal of Experimental Child Psychology*, **78**, 75–83.

Eizenman, D.R., & Bertenthal, B.I. (1998). Infants' perception of object unity in translating and rotating displays. *Developmental Psychology*, **34**, 426–434.

Kohonen, T. (1997). *Self-organizing maps*. Berlin: Springer-Verlag.

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

Munakata, Y., McClelland, J.L., Johnson, M.H., & 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.

Shafer, G., & Mareschal, D. (2001). Modeling infant speech sound discrimination using simple associate networks. *Infancy*, **2**, 7–28.

Shultz, T.R., & Bale, A.C. (in press). Neural network simulation of infant familiarization to artificial sentences; rule-like behavior without explicit rules and variables. *Infancy*.

The modules behind the learning

Gary F. Marcus

Department of Psychology, New York University, USA

Sometimes the most interesting part of a paper turns out to be buried in its appendix. The main part of Mareschal and Johnson's paper is a defense of the idea that the notion of object unity might be learned. But while it may be true that the model that forms the crux of their argument captures some aspects of object unity, careful inspection reveals that learning plays but a minor role in the story that is told here. The appendix reveals that most of the machinery here is actually innate.

When looked at closely, the appendix reveals something straight out of classical cognitive science, something that even Jerry Fodor could love. What really makes the model tick is a set of seven innately given perceptual analyzers – each of which is pre-wired to do a fairly sophisticated bit of domain-specific computation, in a way that is entirely informationally encapsulated. For example, the texture module depicted in Figure 11c counts how many texture units are visible and compares that to how many texture units were visible in the previous time step, yielding a one if the number of texture units has changed, a zero if it has not. Nothing interactive here; and nothing learned either. Similarly, the module depicted in Figure 11a tests for object motion by comparing the current contents of the retina with the contents of a buffer that contains a cached copy of the retina's contents at the previous time step. Again, nothing interactive, nothing learned. But to say that some parts of the architecture are neither interactive nor learned is not an insult – instead, it is a compliment: it is how I think things really are: newborns probably do come to the visual world with a wide array of sophisticated innately wired perceptual analyzers.

Seen in this light, Mareschal and Johnson's overall discussion of the model is underemphasizing a critical component. It is true that the model does learn something, but what it learns is actually the relation between these complex pre-wired perceptual analyzers and an innately given concept of unity: the model learns how an innately given concept of unity correlates with particular cues that are the output of the perceptual analyzers.[1]

A crude way to put it is that the complete model has on the order of a thousand connections,[2] out of which only 18 are adjusted on the basis of experience. Experience is an important component in the organization of the model, but the vast majority of the model's connections are innately wired, and that fact needs to be taken seriously.

[1] In the current model the output units innately represent unity versus non-unity, doing so prior to experience on 'nonambiguous' displays. Of course, as Mareschal and Johnson rightly note, footnote 2, the fact that a perceptual analyzer is functioning at the time of testing doesn't mean that it is hardwired, but to minimize the amount of innate machinery a theorist would need to show a way in which perceptual analyzers themselves could be learned, something that has never been done. A related point is that Mareschal and Johnson's appeal to 'direct perception' is misleading – the world doesn't tell us how many objects are out there – what Mareschal and Johnson are calling direct perception is actually the output of a complex (and here unlearned) computation. (See also discussion in Marcus, 1998.)

[2] There is not enough information given to calculate this number precisely, but I note that each of the seven perceptual analyzers has at least 196 connections, all pre-wired.

Address for correspondence: Department of Psychology, Room 306, 6 Washington Place, New York University, New York 10012, USA; e-mail: gary.marcus@nyu.edu

What I think is being left out of connectionist models of development is the role of *genetics* in helping to sculpt the mind (Marcus 2001a, 2001b). Some of the brain is calibrated or tuned on the basis of experience, but if the brain is anything like the rest of the body, we can expect mechanisms of gene expression to play an important role in brain development. As neuroscientists Lawrence Katz, Michael Weliky and Justin Crowley (2000) recently put it: 'The current emphasis on correlation-based models, which may be appropriate for later plastic changes, could be obscuring the role of intrinsic signals that guide the initial establishment of functional architecture.'

Some of the structure of the brain is learned, but it seems certain that some of it simply grows under genetic guidance, in the absence of experience. We won't understand the exquisite interplay between the innate and the learned until we take both sides of the equation seriously.

References

Katz, L.C., Weliky, M., & Crowley, J.C. (2000). Activity and the development of the visual cortex: new perspectives. In M.S. Gazzaniga (Ed.), *The new cognitive neurosciences* (pp. 199–212). Cambridge, MA: MIT Press.

Marcus, G.F. (1998). Can connectionism save constructivism? *Cognition*, **66**, 153–182.

Marcus, G.F. (2001a). *The algebraic mind: Integrating connectionism and cognitive science*. Cambridge, MA: MIT Press.

Marcus, G.F. (2001b). Plasticity and nativism: towards a resolution of an apparent paradox. In S. Wermter, J. Austin & D. Willshaw (Eds.), *Emergent neural computational architectures based on neuroscience*. New York: Springer-Verlag.

Modeling infants′ perception of object unity: what have we learned?

Yuko Munakata and Jennifer Merva Stedron

Department of Psychology, University of Denver, USA

Mareschal and Johnson have chosen well by exploring infants' perception of object unity through connectionist models. Object unity is a rich domain, with a clear developmental time course established through systematic testing from birth through the first several months of life. Connectionist models provide a powerful tool for exploring the learning mechanisms and environmental input that may contribute to infants' developing abilities to perceive object unity. Mareschal and Johnson's model simulates infants' progression from perceiving partly occluded objects as disjoint to perceiving object unity.

However, for models to advance our understanding in any domain, they must do more than simulate data; they should provide insight into the processes underlying the observed behaviors. In this case, the key questions are: Based on the model, what have we learned about how infants perceive object unity? What have we learned about how this ability develops? The current model may provide a stronger answer to the second question (at a very general level) than to the first. The model develops by learning which environmental cues are most strongly associated with single objects, and then relying more heavily on those cues in perceiving object unity. This general approach seems promising given that infants pick up on regularities in the environment (e.g. Haith, Hazan & Goodman, 1988; Saffran, Aslin & Newport, 1996), so that they might be able to use such information in learning to perceive object unity.

As elaborated below, however, several aspects of the model's processing do not seem to map on to infant processing (see Table 1). These discrepancies pose a challenge to the model's ability to inform us about how infants perceive object unity and how this process develops.

Bias

After the models learn to perceive object unity, they become biased toward unity. That is, they are biased to

Address for correspondence: Department of Psychology, University of Denver, 2155 S. Race St., Denver, CO 80208, USA; e-mail: munakata@du.edu or jstedron@nova.psy.du.edu

Table 1 *Apparent discrepancies between models and infants in the processing of object unity*

Component of perceiving object unity	Models	Infants
Bias	Favor unity	Unlikely
Effect of visual features		
Texture	Hurts	Helps
Motion	Hurts	Helps
Effect of environment	Enriched causes problems with simple events	Unlikely
Target signal	Decays after object partly occluded	No apparent effects of delay on object unity

activate the 'one object' output unit in the absence of any input. Such a bias might suggest that when we open our eyes after closing them, or when we enter a new room, we should be biased to expect to see one object rather than two. It seems unlikely that such a bias subserves infants' (or adults') perception of object unity. This raises the questions of why the models exhibit this bias, and how relevant this solution for perceiving object unity is for infants.

Effect of visual features

Motion and background texture help infants to perceive object unity, but hurt the model's ability to do so. Both of these findings were explained as idiosyncrasies of the network's training environment; motion and changes in background texture happened to be more associated with disjoint objects in the network's environment. The model picked up on this regularity, and so had more difficulty perceiving object unity in the presence of motion and texture. This is a clear explanation of puzzling behavior in the model. However, if training idiosyncrasies are to blame for the model's failure to simulate certain aspects of the behavioral data, how can we interpret the model's successes in simulating the behavioral data? That is, how can we distinguish whether such successes reflect processing mechanisms and environments that are relevant to infants, or whether these, too, result from idiosyncrasies in the training environment of the networks?

Effect of environment

Models 'raised' in relatively enriched learning environments (consisting of eight different events) developed more general knowledge of object unity than models in relatively simple learning environments (consisting of four events). Such effects of environment might be interesting to explore further in connectionist models, given that animals reared in deprived environments show clear behavioral and neural impairments (Greenough, Black & Wallace, 1987; Wiesel & Hubel, 1963). However, the models raised in the relatively enriched learning environments exhibited some very strange behaviors, raising questions about the processes underlying their perception of object unity. For example, these models had great difficulty correctly perceiving unity in Event 1, which was the most simple event (e.g. containing all of the cues that signal unity for infants) and was part of the training set. In contrast, models in the deprived environments showed much better performance with this simple event. Further, the models in the enriched environment successfully perceived object unity in more complex events (e.g. Events 5 and 6) well before perceiving unity in the simple event. This strange pattern of performance suggests that the models in the enriched environment learned to perceive object unity in an anomalous way. Thus, although the global performance measures tell a compelling story (enriched environments support better learning), the detailed patterns of performance in the models raise questions about how meaningful their behaviors are.

Target signal

The models learn object unity via a decaying target signal; after an object is occluded, this decaying target signal specifies that the partially occluded object is a single object. It is not clear where this object unity target signal would come from in infants (in contrast with other backpropagation networks that use continual input from the environment as the training signal; e.g. Elman, 1990; Munakata, McClelland, Johnson & Siegler, 1997). The fact that infants habituate (and thereby demonstrate memory) is tenuous evidence for positing the existence of an activation-based target signal specifying object unity. In fact, the existence of such a target signal, present from the model's 'birth', might be viewed as more in line with nativist than empiricist approaches.

Further, because the target signal decays, it seems likely that even mature models would become less and less confident over time that a partially occluded object is a single object. That is, the models would activate the unity output unit less and less over subsequent occluded time steps. This does not seem to map on to anything in infants (or adults). Both infants and adults can perceive the unity of an occluded object even without first seeing the object fully visible, suggesting that the delay between fully visible and partially occluded views is not a critical factor in humans in the way that it may be in the models.

Conclusions

All models involve simplifications that one can challenge (in terms of what they map on to in the infant, in the environment, etc.). Thus, the evaluation of models should not be viewed as simply a black-or-white process, where good models do not require simplifications but bad models do. Instead, the limitations must always be balanced with the insights a model may provide. In this case, the primary strength of the model seems to lie in its demonstration of the ability to use regularities in the environment to learn to perceive object unity. However, it is already well known that connectionist models can learn statistical regularities from their environments (Rumelhart & McClelland, 1986; O'Reilly & Munakata, 2000). The more useful demonstration would be to show how the model provides insights into the details of how infants perceive object unity and how this ability develops. It is unclear what these insights would be, given the numerous apparent discrepancies between the models and infants in their perception of object unity.

References

Elman, J.L. (1990). Finding structure in time. *Cognitive Science*, **14**, 179–211.

Greenough, W.T., Black, J.E., & Wallace, C.S. (1987). Experience and brain development. *Child Development*, **58**, 539–559.

Haith, M.M., Hazan, C., & Goodman, G. (1988). Expectation and anticipation of dynamic visual events by 3.5-month-old babies. *Child Development*, **59**, 467–479.

Munakata, Y., McClelland, J.L., Johnson, M.H., & Siegler, R. (1997). Rethinking infant knowledge: toward an adaptive process account of successes and failures in object permanence tasks. *Psychological Review*, **104**(4), 686–713.

O'Reilly, R.C., & Munakata, Y. (2000). *Computational explorations in cogntive neuroscience: Understanding the mind by simulating the brain*. Cambridge, MA: MIT Press.

Rumelhart, D.E., & McClelland, J.L. (1986). PDP models and general issues in cognitive science. In D.E. Rumelhart, J.L. McClelland, & PDP Research Group (Eds.), *Parallel distributed processing. Volume 1: Foundations* (Chap. 4, pp. 110–146). Cambridge, MA: MIT Press.

Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical learning by 8-month-old infants. *Science*, **274**, 1926–1928.

Wiesel, T.N., & Hubel, D.H. (1963). Single-cell responses in striate cortex of kittens deprived of vision in one eye. *Journal of Neurophysiology*, **28**, 1029–1040.

Teleology in connectionism

Linda B. Smith

Department of Psychology and the Program in Cognitive Science, Indiana University, Bloomington, IN, USA

Theorists as diverse as Baldwin (1906), Darwin (1877), Piaget (1952) and Werner (1957) viewed development as a process of 'getting better' – as a generative force for the improvement of mankind. In the contemporary literature, development is also seen as more-or-less steady progress toward a goal, toward the adult standard or toward functional and adaptive behavior. We as developmental theorists know where development is going, or where it

Address for correspondence: Department of Psychology and the Program in Cognitive Science, Indiana University, Bloomington, IN 47405, USA; e-mail: smith4@indiana.edu

ought to be going. But does the developing organism know? Does the infant in some real sense represent the endstate? According to nativist-rationalist accounts, they do. This is an honest and direct stance. However, the danger of teleology is that it worms its way into theories meant to oppose the rationalist view.

Mareschal and Johnson's paper is admirable on many grounds. They provide an elegant and insightful review of the data and make a compelling argument that object unity is learned. Most specifically, the conditions under which infants interpret a partly occluded object as a unitary whole change with development. Further, there are multiple co-occurring cues that predict unity – motion, co-motion, common motion, relatability, parallelism, texture and T-junctions. Thus, Mareschal and Johnson argue, by learning these cues, infants could learn to perceive object unity despite occlusion. All that is needed for learning is right there in the everyday experiences of infants. As a whole object in full view moves and becomes partially occluded and then back into full view, the infant can learn which cues correlate with unity and thus come to see partially occluded forms as unitary.

Mareschal and Johnson then take their argument one step further and present a connectionist simulation to show the plausibility of this account. The simulations are revealing about how and in what ways specific cues might be learned. Moreover, the simulations are likely to lead to interesting future experiments. The authors are to be congratulated on this effort; much more formal modeling of developmental phenomena is needed. But, the real value of formal modeling is that the formalisms (unlike explanations couched only in words) are brutally honest. The equations and algorithms leave the theorist no place to hide; they make the mechanism explicit. In the present case, the mechanism is pre-knowledge about the very thing to be learned, object unity. The network knows at the start that object unity is the goal, and it knows what unity is. The learning mechanism is highly constrained and is solely dedicated to learning about the motion patterns that co-occur with objects *already recognized as coherent wholes*. In this way, the model is an implementation version of Kellman and Spelke's (1983, p. 521) rationalist idea that 'humans may begin life with the notion that the environment is composed of things that are coherent, that move independently of one another, and that tend to persist . . .'.

An implementation on the rationalist side

Where do our rich understandings of objects, causality, space, number, time come from? Rationalists look at the diversity of human knowledge and the certainty with which children acquire it and conclude that the diversity is there to begin with, that there are special mechanisms, operating in accord with specific principles, dedicated to specific domains (e.g. Spelke, Breinlinger, Macomber & Jacobson, 1992). Two main ideas underlie this rationalist view (see Fodor, 1983):

1. What needs to be acquired is unique to each domain. Thus, the mechanisms that guide learning in each domain must be domain specific, pertinent to the specific content and task in that domain.
2. One cannot get something from nothing, and thus nothing truly new can be learned. What is to be learned must be pre-specified.

Mareschal and Johnson's model adheres to these two rationalist ideas.

First, the features to which the model is sensitive are all relevant to unity. Motion, co-motion, common motion, relatability, parallelism, texture and T-junctions are pre-selected and bundled into this one dedicated device. There are no features such as distance, or number, or velocity, or shape malformation, or contact that are not relevant to unity but are relevant to learning causality, space, number or object kind. Thus, the model accords with the first main idea of rationalism – a dedicated mechanism that is highly constrained and prebuilt to learn just one thing.

Second, the device knows what unity is at the start; all it learns is the specific motion patterns during occlusion that co-occur with a unitary unoccluded object. Moreover, the device can only learn whatever cues predict unity; it cannot learn anything else. This derives from the very nature of the learning algorithm. The network uses an error reduction learning algorithm (backpropagation). This algorithm requires that the learner *know* what is to be learned. This algorithm works by minimizing the distance between the actual output on any trial and a target output. In this case, the target output is the perception of the unity of the object – its coherence as a whole – when unoccluded. This network does not learn

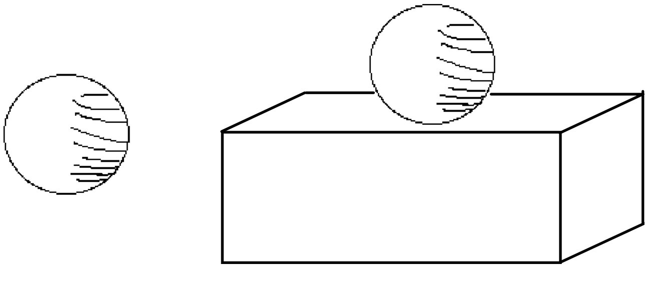


Figure 1 *Young infants fail to recognize a ball as an individual entity when it is placed on top of another object.*

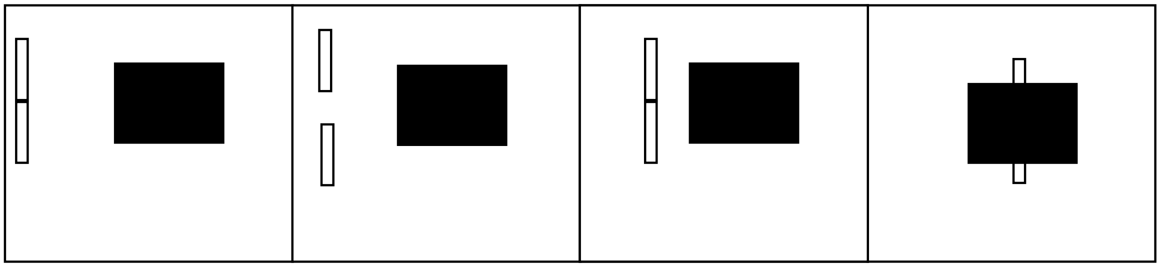


Figure 2 *A series of events showing an object composed of two parts that then moves as a whole behind a screen.*

to perceive unity. It is given the recognition of unity, the value that this matters, all the potentially relevant features (and no irrelevant ones) and the goal of predicting the unity. The target – the perception of unity – defines what is to be learned and does all the work. With this learning algorithm, the network knows what needs to be learned before it does any learning. In this way, the mechanism of development *is* the evaluation of how close development is to a pre-specified endstate. Thus, the model adheres to the second rationalist tenet; nothing really new is learned.

Unity is not all there is

Piaget (1952) observed that a young baby's visual tracking of an object or an older baby's reach for an object could be disrupted by placing the object on top of or behind another object. According to Piaget, an object such as the ball in Figure 1 ceases to be perceived as a separate object when it is placed on top of another. This model as it stands would make the same prediction, since it is designed to perceive unity; without motion, the ball and block are one. But unlike children, Maraschal and Johnson's model could not eventually learn to perceive these stationary objects as separate based on their configural properties alone. Indeed, experiences in which separate objects are temporarily put together seem likely to cause real problems for this device. Consider the sequence of events presented in Figure 2. If these were presented to the network during the learning phase, what would the network learn? The first step in the input should specify a unitary whole object and thus as configured, the network should learn that this pattern of motion also predicts a unitary whole. With unity as the pre-specified target of learning, there is nothing else to learn.

If we follow the approach of Mareschal and Johnson's model, the only way to learn other things is to build lots of other little dedicated devices prebuilt to learn what needs to be learned. The problem is that babies have lots to learn – about objects, about motions, about number, about causality. If we follow this approach, we will end up where the rationalists did, with many, many innate ideas.

Any way out?

Can we get teleology out of developmental mechanism? Apparently, it is not easy. But it might be easier if we dropped the idea that we know what development is for and that we know where it is going. At the very least, connectionist modelers might avoid error driven learning algorithms that pre-specify what is to be learned.

References

Baldwin, J.M. (1906). *Social and ethical interpretations of mental development*. New York: Macmillan.

Darwin, C. (1877). A biographical sketch of an infant. *Mind*, **2**, 285–294.

Fodor, J.A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Kellman, P.J., & Spelke, E.S. (1983). Perception of partly occluded objects in infancy. *Cognitive Psychology*, **15**, 483–524.

Piaget, J. (1952). *The origins of intelligence in children*. New York: International Universities Press.

Spelke, E.S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review*, **99**, 605–632.

Werner, H. (1957). The concept of development from a comparative and organismic point of view. In D.B. Harris (Ed.), *The concept of development* (pp. 125–148). Minneapolis: University of Minnesota Press.

RESPONSE

Of models and mechanisms: a reply to commentators

Denis Mareschal[1] and Scott P. Johnson[2]

1. *Centre for Brain and Cognitive Development, School of Psychology, Birkbeck College, University of London, UK*
2. *Department of Psychology, Cornell University, USA*

We were pleased to see that all commentators agreed with us on one issue. All four authors argued that computational modeling is a fertile way to push forward debates in cognitive development. Computational models are explicit instantiations of information processing theories. They describe exactly what is assumed to be in the environment, what is assumed to be 'built in', and some of the mechanisms by which development can occur. It is precisely because they are 'brutally honest' (Smith) that they can push debates forward. By laying our cards on the table and inviting commentators to specify what they believe to be wrong with the assumptions of our model, we are also requiring *them* to be precise about what might constitute the mechanisms of development of unity perception in infants.

Model building is an iterative process. At each iteration, the model is improved to take account of more data, or perhaps to correct some invalid assumptions. Model development is also a means of fostering dialogue. Building a model invites those with different views to participate by expressing their dissatisfaction with a current implementation. If the criticism is sufficiently precise, it becomes an agreed target towards which both research groups can work, the presumption being that if a revised model can overcome the obstacle or limitation highlighted, then the two views will converge.

We were equally happy to see that the four commentators have engaged with us in this dialogue by highlighting aspects of the model that they endorse, as well as aspects of the model with which they would find fault. Of course, models are necessarily approximations (Rummelhart & McClelland, 1986; Mareschal, 2001), and building a model is not the same as building an infant! As a result, we agree with some of the points made, and disagree with others. In what follows we will outline how we might address each of the main concerns.

Nativism through the back door (Marcus, Smith and Munakata & Stedron): Is the concept of unity somehow innately specified in the model?

No. To be precise, the concept of unity as tested in the ambiguous displays depicted in Figure 1 (Mareschal & Johnson, this volume) is not. The model has a pre-existing feature detector that allows it to differentiate between the presence of one or two objects when all portions of the object are visible. Although we have labeled this as a unity input, it might be better described as a form of low-level subitization (i.e. discriminating one from two objects). This ability is presupposed in all empirical tests of infants' unity perception, because it is required to discriminate between test stimuli (b) and (c) in Figure 1. If infants were not able to discriminate these stimuli, then the empirical results would lose their meaning. As noted (Mareschal & Johnson, this volume), even neonates are able to make this distinction (Slater, Johnson, Brown & Badenoch, 1996).

The models learn, ultimately, how to make this discrimination in a completely ambiguous event, an event in which the low-level unity feature detector is unable to compute whether there is one or two objects in the event because of the partial occlusion. In these cases, the model learns to generate a prediction about unity on the basis of other available cues. This mediated response is not wired in and emerges through interactions with the environment. Moreover, it allows the network to respond

Address for correspondence: Denis Mareschal, Centre for Brain and Cognitive Development, School of Psychology, Birkbeck College, Malet Street, London WCIE 7HX, UK; e-mail: d.mareschal@bbk.ac.uk or Scott P. Johnson, Department of Psychology, Uris Hall, Cornell University, Ithaca, NY, 14853, USA; e-mail: sj75@cornell.edu

to ambiguous events in a manner analogous to human infants.

Modeling unity (Cohen & Chaput, Marcus): What exactly are we trying to model?

We are trying to model infants' perception of partly occluded objects. The assumption underlying extant theories of unity perception is that these percepts, whether veridical (i.e. adultlike) or not, are formed outside the laboratory. Cohen and Chaput are correct in pointing out that the habituation paradigm can tap either learning during habituation itself, or some pre-existing perceptual or cognitive capacity. Habituation is a tool for assessing these capacities, a tool that is useful for probing infants' responses to the stimuli we present. But it is hardly necessary for us to model habituation to address the questions we are asking. We assume that during development, infants are exposed to both fully visible and partly occluded objects as well as the transitions between these views, and we believe this assumption is captured well in the models.

We are not trying to model the development of the visual system as a whole. Marcus appears to find fault with our 'detection, then utilization' hypothesis (although this is not stated explicitly), in that our models begin with fully functional, encapsulated modules that each process one attribute of the visual scene. We see this critique as more of a restatement of our approach than as a problem. The models begin the process of learning unity with much of the hardware in place, as does the human neonate, who is born with a full complement of cortical neurons (Rakic, 2000). Frankly, we were disappointed with the simplistic presentation of 'genetics' as *the* principal sculptor of brain structure, and we reject outright the notion that there is some sort of 'equation' or continuum between 'genetic' and 'learned'. This is precisely the line of thinking that we hope to expel from debates about the origins of knowledge, because it buys us nothing. We want to know about *mechanisms*, no matter what they may be. Several mechanisms are discussed in the chapter Marcus cites that contribute to visual system development, and these are functional before the onset of visual experience (Katz, Weliky & Crowley, 2000). To describe these mechanisms as 'genetic', however, is a mischaracterization. Indeed, they are paradigmatic examples of the exquisitely intricate interplay between gene expression, molecular markers (transcription factors and axon guidance molecules), plasticity, experience, and so on, that mark all development. For example, Katz *et al.* describe spontaneous activity of retinal cells during fetal development that provides correlated structure with which downstream areas (such as LGN in the thalamus and visual cortex) may derive patterned connectivities and from which orientation selectivity and cortical topographic mapping may emerge (cf. Pallas, 2001; Wong, 1999). The crude mapping, present upon eye opening, is then fine-tuned with visual experience. There is no obvious place here where the role of 'genetics' begins and ends.

We are not trying to model all aspects of cognitive development, such as understanding of 'causality, space, time and number' (Smith). We fully concur with Smith's contentions that our model learns 'just one thing' and that 'unity is not all there is'. Our rejoinder to this criticism is that some narrowness of focus is a feature of *any* empirical or modeling investigation. We would go farther, in fact, in noting that the mechanism by which our models learned is only *one* possible route to veridical object perception, and is certainly not the entire story (see Johnson, 2001).

Error-driven supervised learning (Cohen & Chaput, Smith): Is this really a supervised network masquerading as an unsupervised network?

No. It is important to understand the distinction between *backpropagation networks* and *supervised learning*. In supervised learning, an external agent provides the network with an example of a correct response (e.g. a category label or a past-tense morpheme). Unsupervised networks are designed to self-organize in order to reach a consistent response in the absence of an external *agent* supplying a teaching signal. They come into equilibrium with their environment according to some constraints determined by their wiring and/or the specific learning algorithm.

Backpropagation is one kind of algorithm for updating weights in the network. It functions by minimizing the difference between the network's output response and some other signal. In many applications, that other signal is the target information provided by an external agent. Thus, backpropagation is often used in a supervised training context. However, this does not necessarily have to be the case, and it is not the case with the present model. *Autoencoders* are examples of networks that use backpropagation but do not rely on an external training agent (Mareschal, French & Quinn, 2000; Japkowicz, 2001). These networks try to minimize the difference between their output and the original input from the environment. That is, they are designed to reproduce, on the output, exactly the same input features that they have just encountered. As a result, they develop internal representations that capture the key feature variation in the environment.

Autoencoders are sometimes called 'self-supervised' systems, and it may be easier to think of our networks as 'self-supervised' rather than self-organizing. They are engaged in a process of self-organization (driven by the autoencoding task) that causes the emergence of internal representations that maximize the network's ability to predict the features of its particular environment. Although not a full autoencoder,[1] it is self-supervised. The target signal used to train the weights with backpropagation is obtained directly from the environment (as in autoencoder models) or from a decaying memory trace that is internal to the system.[2] *There is no external agent telling the network whether a particular event is unified or not*. Hence, the networks receive no more information than do infants in real-world situations of learning object properties.

The inductive bias (Munakata & Stedron, Cohen & Chaput): Do our models instantiate an inductive bias that in and of itself determines correct performance?

An 'inductive bias' is a technical term from the formal learning literatures (Mitchell, 1997), and refers to any computational bias that allows the system to select preferentially one from a number of equally probable options. Both the test stimuli (depicted in Figure 1b and 1c of Mareschal & Johnson, this volume), for example, are equally likely sources of the ambiguous habituation event (Figure 1a). There is nothing in the occlusion display itself that allows an observer to choose between them rationally. The fact that infants and adults consistently choose one or the other, however, is evidence that an inductive bias exists in the computations of their perceptual systems. Importantly, that bias appears to change with age. The bias in neonates leads them to perceive the partly occluded surface in a rod-and-box display as arising from two objects (i.e. the rod is perceived as a disjoint object). The bias in 4-month-olds (and adults), in contrast, leads them to perceive the partly occluded surface as arising from a single unified object.

Note that we are inferring the presence of an inductive bias from observed behavior only. We have said nothing about how that bias is implemented or materialized in the infant's visual system. Similarly, when we talk about an inductive bias in the networks' behavior, we are referring to their consistent response to an ambiguous event but we remain silent on the precise mechanisms underlying this bias.

It is important not to confuse the 'inductive bias' of the system with the 'bias nodes' in the network. The bias nodes are a way of implementing a tuneable firing threshold in the output and hidden units. This threshold may or may not be involved in implementing the observed inductive bias but is not the same thing. Early in development the 'bias' towards seeing the ambiguous event as composed of two disjoint objects may be due to the greater proportion of unified events in the networks' environment, but this is not the case later in training. The 'inductive bias' shifts but the environment remains constant. Hence, we have not surreptitiously built in the bias through our training environment.

The higher resting threshold of unity output unit in 'adult' networks translates into the prediction that adults and 4-month-olds would be slightly quicker to identify a single unified object than two disjoint objects, but that the converse would hold for neonates. Of course, a higher resting threshold for the unity node does not imply a change in the adult's (nor the 4-month-old's) percept in any way because as soon as a stimulus appears, other perceptual cues (e.g. T-junctions) come on line and contribute accordingly to the computation of the percept. The final result, then, is a consistent tendency to see unity, or a single object, despite partial occlusion. This captures very well the tendency of human adults to perceive partly occluded surfaces as arising from unified objects, but it does not imply that we would be biased toward seeing any specific number of objects when we first encounter a scene (as suggested by Munakata & Stedron).

Match between infant and model behavior and experience (Cohen & Chaput, Munakata & Stedron, Smith): How effectively do our models capture the transition across early infancy toward veridical perception of object unity, and how closely does our artificial environment resemble that of infants?

Although the model generally fits the developmental profile observed in infants, a number of commentators

[1] In fact, these networks can be seen as partial autoencoders. Their task is to reproduce on the output units the initial low-level feature description produced by the collective outputs of the perceptual modules. However, rather than modeling the complete autoencoding of all features, we have only modeled how the autoencoding of a single feature (the unity feature) would be computed. This is a legitimate approach because all the output features in an autoencoder are computed independently from the same hidden unit representation. An interesting direction to follow up this work would be to see how full auto association would effect the network's performance on the unity task and how the internal representations across the hidden units would change.

[2] One can think of the decaying target trace as a short-term iconic visual memory. It could be implemented using a parallel bank of perceptual feature units with decaying activation.

have pointed out that there are significant ways in which it also differs from infant performance. We argued in the target article that these differences reflect the divergence between the content of our schematic environments and the infant's far richer experiences in the real world. Responses consistent with those interpreted as evidence of a concept of unity can be acquired by an associative learning system. Through interactions with the environment, the system learns which combinations of perceptual cues predict unity in an ambiguous context. The networks' training environment is overly simplified, but we feel it is representative of experience infants may have.

Notably, the complexity of the environment is important in determining the generality of the unity response acquired. Training in the more complex environment slows down the acquisition of a veridical unity response on the full cue event (as compared to the model's performance with a limited environment), but it greatly improves performance on unity problems as a whole (cf. Figures 7 and 8). This apparent delay is caused by the fact that in the full cue event, the networks had to learn to deal with a large number of cues that sometimes were associated with unity, and sometimes not. (Having more cues does not necessarily mean that a problem is easier if some of those cues are not reliable indicators!) A straightforward prediction from these results is that infants will learn to respond appropriately to events that contain a very reliable indicator (such as the lack of common motion in events 15–26, a predictor of disjoint objects) well before they are able to perform correctly on the ambiguous full cue event (i.e. events 1 and 2). Similarly, motion, in some cases, appears to have a detrimental effect on the perception of object unity. This finding is somewhat counterintuitive given the importance of many kinds of rigid motion to infants' unity perception (Kellman, 1996). However, not all rigid motions reliably specify unity to infants: as noted by Cohen and Chaput, Eizenman and Bertenthal (1998) reported that infants fail to perceive unity in a stimulus in which a rod is seen to oscillate in the frontoparallel plane until around 6 months, in contrast to much earlier unity perception when the rod undergoes lateral translation (2 months; Johnson & Aslin, 1995). In other contexts, motion can have both an inhibitory and facilitory effect on infants' object perception (see Burnham, 1987, for a review). Our models do not do justice to the complexities of real-world motion of objects, of course, and much remains to be explored, both in modeling and empirical testing of the role of motion in the emergence of object perception.

Cohen and Chaput suggest that our environment might be 'giving away the answer' to the question of unity percepts by training the models in very specific environments, and we acknowledge that this is a fundamental challenge. An important point growing from this issue is the nature of the *infant's* environment as he or she is able to perceive it. There are two serious impediments to a full understanding of how experience can contribute to infants' object knowledge. First, we know little about real-world scenes in terms of the frequency with which objects are seen in their entireties, the extent to which they are occluded when behind nearer surfaces, and the transitions between these two circumstances. There are indications that the statistics of natural scenes have a powerful influence on cortical visual development, in terms of orientation selectivity (Coppola, Purves, McCoy & Purves, 1998; Olshausen & Field, 1996). An investigation of the probabilistic nature of occlusion in the optic array, likewise, would prove invaluable to our understanding of infant cognitive development. Second, we cannot assume that the infant's experience in the real world will match an adult's, because of obvious limitations in basic visual function (such as acuity) and exploratory action (such as eye movements, prehension and locomotion). Such limitations will hamper the infant's ability to gain information in an efficient way (Campos, Anderson, Barbu-Roth, Hubbard, Hertenstein & Witherington, 2000; Johnson & Johnson, 2001).

What, indeed, have we learned? (Munakata & Stedron): What do our models tell us about infant development?

Munakata and Stedron take the models to task for failing to provide 'insights into the details of how infants perceive object unity and how this ability develops', implying that we have done little more than to document the capacity of connectionist models to learn statistical regularities. We would reply that infants can learn statistical regularities too (Saffran, Aslin & Newport, 1996), even as young as 2-month-olds (Kirkham, Slemmer & Johnson, in press); does knowing this mean that we have solved all the problems of cognitive development?

The contributions of our models can be stated in Munakata and Stedron's own terms, which were expressed elsewhere in a different context (Munakata & Stedron, 2001). First, our models allow us to 'compare competing theories' (p. 166) by documenting the plausibility of a learning account of unity perception, as opposed to a core principles account (e.g. Spelke & Van de Walle, 1993), thereby addressing a longstanding and fundamental debate. Second, our models adhere to the Munakata and Stedron (2001) maxim that 'simple is good' (p. 167) by demonstrating how relatively complex object knowledge can arise out of simpler perceptual mechanisms that are likely to be functional in the youngest infants. Third, models 'can fail and provide insights when they do' (p. 167), an apt descriptor of our models' tendency to get tripped up by texture information. It is probable that infants use texture information in a different way than did our models (i.e. as depth information; see Johnson & Aslin, 1996), and the differences between infants and models in this respect remain a fruitful question for future research. Fourth, the models 'have something to say about change as well as origins' (p. 164). Figure 10 in Mareschal and Johnson (this volume) depicts the development of cue use in terms of connection strengths between perceptual modules and hidden units. As discussed in the section 'Learning cue associations', a clear picture emerges of increasing veridicality of object perception following reliance on multiple cues. Little is known about similar processes in human infants, but the idea is highly plausible, and testing of this hypothesis is under way. Fifth, and finally, Munakata and Stedron (2001) provided the astute observation that 'we will make the most progress by specifying alternative models that build on existing strengths and begin to address limitations' (p. 170). We share this position, noting that models are an invitation to a dialogue. All models are approximations, and only approximations. Our models represent but one stage in the long effort to understand mechanisms of development and the origins of object knowledge, and we look forward to further empirical efforts from the commentators and others as we approach this goal.

References

Burkhalter, A., Bernardo, K.L., & Charles, V. (1993). Development of local circuits in human visual cortex. *Journal of Neuroscience*, **13**, 1916–1931.

Burnham, D.F. (1987). The role of movement in object perception by infants. In B.E. McKenzie & R.H. Day (Eds.), *Perceptual development in early infancy: Problems and issues* (pp. 143–172). Hillsdale, NJ: LEA.

Campos, J.J., Anderson, D.I., Barbu-Roth, M.A., Hubbard, E.M., Hertenstein, M.J., & Witherington, D. (2000). Travel broadens the mind. *Infancy*, **1**, 149–219.

Coppola, D.M., Purves, H.R., McCoy, A.N., & Purves, D. (1998). The distribution of oriented contours in the real world. *Proceedings of the National Academy of Sciences*, **95**, 4002–4006.

Eizenman, D.R., & Bertenthal, B.I. (1998). Infants' perception of object unity in translating and rotating displays. *Developmental Psychology*, **34**, 426–434.

Edelman, S. (1997). Computational theories of object recognition. *Trends in Cognitive Sciences*, **1**, 296–304.

Japkowicz, N. (2001) Supervised and unsupervised binary learning by feedforward neural networks. *Machine Learning*, **42**, 97–122.

Johnson, S.P. (2001). Visual development in human infants: binding features, surfaces, and objects. *Visual Cognition*, **8**, 565–578.

Johnson, S.P., & Aslin, R.N. (1996). Perception of object unity in young infants: the roles of motion, depth, and orientation. *Cognitive Development*, **11**, 161–180.

Johnson, S.P., & Johnson, K.L. (2001). Early perception-action coupling: eye movements and the development of object perception. *Infant Behavior and Development*, **23**, 461–483.

Katz, L.C., Weliky, M., & Crowley, J.C. (2000). Activity and the development of the visual cortex: new perspectives. In M.S. Gazzaniga (Ed.), *The new cognitive neurosciences* (pp. 199–212). Cambridge, MA: MIT Press.

Kellman, P.J. (1996). The origins of object perception. In E. Carterette & M. Friedman (Series Eds.) & R. Gelman & T. Au (Vol. Eds.), *Handbook of perception and cognition: Perceptual and cognitive development* (2nd edn., pp. 3–48). San Diego: Academic Press.

Kirkham, N.Z., Slemmer, J.A., & Johnson, S.P. (in press). Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition.*

Mareschal, D. (2001). Connectionist methods in infancy research. In J. Fagen & H. Hayne (Eds.), *Progress in infancy research*, *Vol. 2* (pp. 71–119). Mahwah, NJ: Erlbaum.

Mareschal, D., French, R.M., & Quinn, P.C. (2000). A connectionist account of asymmetric category learning in early infancy. *Developmental Psychology*, **36**, 635–645.

Mitchell, T.M. (1997). *Machine learning*. New York: McGraw-Hill.

Munakata, Y., & Stedron, J.M. (2001). Neural network models of cognitive development. In C.A. Nelson & M. Luciana (Eds.), *Handbook of developmental cognitive neuroscience* (pp. 159–171). Cambridge, MA: MIT Press.

Olshausen, B.A., & Field, D.J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, **381**, 607–609.

Pallas, S.L. (2001). Intrinsic and extrinsic factors that shape neocortical specification. *Trends in Neurosciences*, **24**, 417–423.

Rakic, P. (2000). Setting the stage for cognition: genesis of the primate cerebral cortex. In M.S. Gazzaniga (Ed.), *The new cognitive neurosciences* (pp. 7–21). Cambridge, MA: MIT Press.

Rumelhart D.E., & McClelland J.L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1.* Cambridge, MA: MIT Press.

Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical learning by 8-month-old infants. *Science*, **274**, 1926–1928.

Slater, A., Johnson, S.P., Brown, E., & Badenoch, M. (1996). Newborn infants' perception of partly occluded objects. *Infant Behavior and Development*, **19**, 145–148.

Spelke, E.S., & Van de Walle, G. (1993). Perceiving and reasoning about objects: insights from infants. In N. Eilan, R.A. McCarthy & B. Brewer (Eds.), *Spatial representation: Problems in philosophy and psychology* (pp. 132–161). Oxford: Blackwell.

Wong, R.O.L. (1999). Retinal waves and visual system development. *Annual Review of Neuroscience*, **22**, 29–47.