The Development of Embodied Cognition: Six Lessons from Babies

Linda Smith
Psychology Department
Indiana University
Bloomington, IN 47405
smith4@Indiana.edu

Michael Gasser
Computer Science Department
Indiana University
Bloomington, IN 47405
gasser@Indiana.edu

Abstract The embodiment hypothesis is the idea that intelligence emerges in the interaction of an agent with an environment and as a result of sensorimotor activity. We offer six lessons for *developing* embodied intelligent agents suggested by research in developmental psychology. We argue that starting as a baby grounded in a physical, social, and linguistic world is crucial to the development of the flexible and inventive intelligence that characterizes humankind.

Keywords
Development, cognition, language, embodiment, motor control

1 Introduction

Traditional theories of intelligence concentrated on symbolic reasoning, paying little attention to the body and to the ways intelligence is affected by and affects the physical world. More recently, there has been a shift toward ideas of embodiment. The central idea behind the embodiment hypothesis is that intelligence emerges in the interaction of an agent with an environment and as a result of sensorimotor activity. This view stands in opposition to more traditional notions of internal representation and computation and in general has had little to say about symbols, symbolic reasoning, and language. In this article we offer six lessons for *developing* embodied intelligent agents suggested by research in developmental psychology. We argue that starting as a baby grounded in a physical, social, and linguistic world is crucial to the development of the flexible and inventive intelligence that characterizes humankind. In preview, the six lessons are these:

1. Babies' experience of the world is profoundly multimodal. We propose that multiple overlapping and time-locked sensory systems enable the developing system to educate itself—without defined external tasks or teachers—just by perceiving and acting in the world.
2. Babies develop incrementally, and they are not smart at the start. We propose that their initial prematurity *and the particular path* they take to development are crucial to their eventual outpacing of the world's smartest AI programs.
3. Babies live in a physical world, full of rich regularities that organize perception, action, and ultimately thought. The intelligence of babies resides not just inside themselves but is distributed across their interactions and experiences in the physical world. The physical world serves to bootstrap higher mental functions.
4. Babies explore—they move and act in highly variable and playful ways that are not goal-oriented and are seemingly random. In doing so, they discover new problems and new solutions. Exploration makes intelligence open-ended and inventive.

5. Babies act and learn in a social world in which more mature partners guide learning and add supporting structures to that learning.
6. Babies learn a language, a shared communicative system that is symbolic. And this changes everything, enabling children to form even higher-level and more abstract distinctions.

These ideas are not without precedent in the robotics and artificial intelligence literature. For example, Brooks et al. [6] and Pfeiffer and Scheier [33] have demonstrated how solutions may fall out of physical embodiment. Breazeal [5] and the Kismet project are beginning to explore how social interactions can bootstrap learning. And exploration is a key idea in machine learning, especially reinforcment learning [50]. The lessons to be learned from human development, however, are not fully mined. Greater interaction between those who study developmental processes in children and those who attempt to create artificial devices that develop through their interactions in the world would be beneficial to both sets of researchers. Accordingly, we offer these lessons from the perspective of researchers who study how babies *become* smart.

2 Six Lessons

2.1 Lesson 1: Be Multimodal

People make contact with the physical world through a vast array of sensory systems — vision, audition, touch, smell, proprioception, balance. Why so many? The answer lies in the concept of *degeneracy* [15]. The notion of degeneracy in neural structure means that any single function can be carried out by more than one configuration of neural signals and that different neural clusters also participate in a number of different functions. Degeneracy creates redundancy such that the system functions even with the loss of one component. For example, because we encounter space through sight, sound, movement, touch, and even smell, we can know space even if we lack one modality. Being blind, for example, does not wipe out spatial concepts; instead, as studies of blind children show [25], comparable spatial concepts can be developed through different clusters of modalities.

Degeneracy also means that sensory systems can educate each other, without an external teacher. Careful observers of infants have long noted that they spend literally hours watching their own actions [34, 7] — holding their hands in front of their faces, watching as they turn them back and forth, and, some months later, intently watching as they squeeze and release a cloth. This second characteristic of multimodality is what Edelman [15] calls *reentry*, the explicit interrelating of multiple simultaneous representations across modalities. For example, when a person experiences an apple — and immediately characterizes it as such — the experience is visual, but also invokes the smell of the apple, its taste, its feel, its heft, and a constellation of sensations and movements associated with various actions on the apple. Importantly, these multimodal experiences are time-locked and correlated.

Changes in the way the hand feels when it moves the apple are time-locked with the changes one sees as the apple is moved. The time-locked correlations create a powerful learning mechanism, as illustrated in Figure 1, which shows four related mappings. One map is between the physical properties of the apple and the neuronal activity in the visual system. Another map is between the physical properties of the apple and neuronal activity in the haptic system. The third and fourth maps are what Edelman calls the reentrant maps: Activity in the visual system is mapped to the haptic system, and activity in the haptic system is mapped to the visual system. Thus the two independent mappings of the stimulus — the sight and the feel — provide qualitatively different glosses on the world, and by being correlated in real time, they educate each other. At the same time, the visual system is activated by time-varying changes in shading and texture and collinear movement of points on the apple, and the haptic system is activated by time-locked changes in pressures and textures. At every step in real time, the activities in each of these heterogeneous processes are mapped to each other, enabling the system in its own activity to discover higher-order regularities that transcend particular modalities.

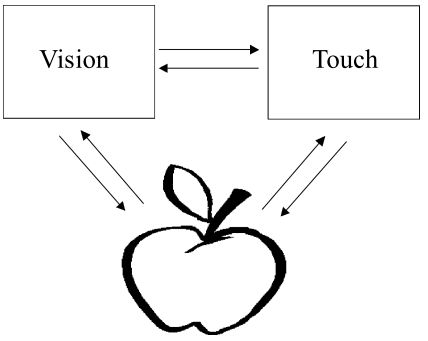

Figure 1. Illustration of the time-locked mappings of two sensory systems to the events in the world and to each other. Because visual and haptic systems actively collect information — by moving hands, by moving eyes — the arrows connecting these systems to each other also can serve as teaching signals for each other.

One clear demonstration of the power of this idea comes from a study of how babies come to understand transparency. Transparency is a problematic concept; think of birds who harm themselves by trying to fly through windows. Transparency is a problem because correlations between visual cues and the haptic cues that characterize most of our encounters with the world do not work in this case. So babies, like birds, are confused by transparency. In one study, Diamond [13] presented infants with toys hidden under boxes such that there was an opening on one side, as illustrated in Figure 2. These boxes were either opaque—hiding the toy—or transparent so that the infants could see the toy under the box. The key result is that 9-month-old infants are better able to retrieve the toy from the opaque than from the transparent container. The problem with the transparent container is that infants attempt to reach for the toy directly, through the transparent surface, rather than searching for and finding the opening.

Infants readily solve this problem, however, if they are given experience with transparent containers. Titzer, Thelen, and Smith [55] gave 8-month-old babies a set of either opaque or transparent buckets to play with at home. Parents were given no instructions other than to put these containers in the toy box, making them available to the infants during play. The infants were then tested in Diamond's task when they were 9 months old. The babies who had been given opaque containers failed to retrieve objects from transparent ones just as in the original Diamond study. However, infants who had played with the transparent containers sought out and rapidly found the openings and retrieved the object from the transparent boxes.

Why? These babies in their play with the containers—in the interrelation of seeing and touching—had learned to recognize the subtle visual cues that distinguish solid transparent surfaces from no surface whatsoever and had learned that surfaces with the visual properties of transparency are solid. The haptic cues from touching the transparent surfaces educated vision, and vision educated reaching and touch, enabling infants to find the openings in transparent containers. These results show how infants' multimodal experiences in the world create knowledge—about openings, object retrieval, and transparent surfaces.

Recent experimental studies of human cognition suggest that many concepts and processes may be inherently multimodal in ways that fit well with Edelman's idea of reentrance [3, 16, 24]. One line

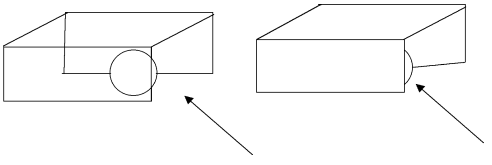

Figure 2. A toy (ball) hidden under a transparent box and an opaque box in the Diamond task. The opening is indicated by the arrow.

of evidence for this conclusion is that even in tasks meant to be explicitly unimodal, multiple modalities contribute to performance. For example, *visual* object recognition appears to automatically activate the actions associated with the object. In one study, adults were shown a picture of a water pitcher such as that illustrated in Figure 3. The task was simple: to indicate by pressing a button whether the object was a pitcher ("yes") or it was not ("no"). Response time was the dependent measure. This is a purely visual object recognition task. Yet the participants were much faster at recognizing the object if the button pressed to indicate the "yes" response was on the same side as the pitcher's handle, as if seeing the handle primed (and readied) the motor response of reaching to that side. Similar results have been reported with a wide variety of objects and in tasks using several different methods. In general, people are faster in visual recognition tasks when the response to be made is compatible with a real action on the object. These results tell us that visual recognition is of a piece with, in the same internal language as, action. This is how it must be under the idea of reentrant mappings, where visual recognition is built out of and educated by its time-locked connections with actions on objects.

2.2 Lesson 2: Be Incremental

Traditionally, both machine learning and human learning have concentrated on non-incremental learning tasks, tasks in which the entire training set is fixed at the start of learning and then is either presented in its entirety or randomly sampled. This is not, however, the way children encounter the world. The experiences of a 3-month-old are very different from (and much more constrained than) the experiences of a 1-year-old, whose experiences, in turn, are very different from those of a 2-year-old. All indications are that these systematic changes in the input, in the range and kind of experiences, matter—that, in fact, they determine the developmental outcome.

Infants' early experiences are strongly ordered by the development of sensory systems and movement systems. At birth, audition and vision are online, but vision is limited by the infant's ability to focus. Nonetheless, shortly after birth, infants look in the direction of sound [31, 57]. Incrementally, over the next few months, subtler and subtler sound properties in relation to visual events begin to take control of visual attention, so that infants *look* at the visual event that matches what they *hear*. For example, given two visual displays of bouncing balls, 4-month-olds look at the displays that are in temporal synchrony with the sound of a bouncing ball [49]. This coupling of hearing and looking organizes infants' attention and thus what they learn. Indeed, children without audition, deaf children, show altered and more disorganized visual attention [47].

Infants' coordination of looking and listening is a form of the reentrant mappings and multimodal learning highlighted under lesson 1. But the important point for lesson 2 is that these correlations *do not stay the same over developmental time*. After looking at and listening to the world for 3 or 4 months, infants begin to reach for objects, and the multimodal correlations change. Once infants can reach, they can *provide themselves* with new multimodal experiences involving vision, haptic

Figure 3. Illustration of the Tucker-Ellis task. On each trial, the participant is shown one pitcher and is asked to answer as rapidly as possible the question: "Is this a pitcher?" On some trials the pitcher's handle is on the left; on some trials it is on the right. Half the participants answer "yes" by pressing a button on the right and half by pressing a button on the left. Participants are faster when the handle is on the same side as the "yes" response.

exploration, proprioceptive input from self-movement, and audition as the contacted objects squeak, rattle, or squeal. After weeks and months of living in this new multimodal venue of sitting, looking, listening, reaching, and manipulating objects, infants' experiences—and the correlations available to them—again change radically, as they begin to crawl and then to stand up and walk. Self-locomotion changes the nature of the visual and auditory input even more dramatically, and the evidence suggests that it also profoundly changes infants' cognitive development.

We review one piece of evidence for this idea that dramatic shifts in the input—shifts that result from changes in the infants' own behavior—cause equally dramatic shifts in cognitive development. Our example involves one of the best-studied tasks of infant cognition, the so-called object-concept or *A*-not-*B* task [48, 51]. Piaget devised this task to assess when infants understand that objects persist in time and space independent of one's own actions on them. In this task, illustrated in Figure 4, the experimenter hides a tantalizing toy under a lid at location *A*. A 3–5s delay is imposed before the infant is allowed to search. Typically infants reach correctly to the hiding location *A* and find the hidden toy. This *A*-location trial is repeated several times. Then, there is the critical switch trial: The experimenter hides the object at a new location, *B*. A brief delay is again imposed, and then the infant is allowed to reach. Infants 8–10 months of age make a curious "error." They reach, not to where they saw the object disappear, but back to *A*, where they had found the object previously. This "*A*-not-*B*" error is especially compelling in that it is tightly linked to a highly circumscribed developmental period; infants older than 12 months search correctly on the critical B trials. Why this dramatic shift in search behavior between 8 and 12 months of age?

The shift appears to be tightly tied to self-locomotion, which also emerges in this same period. Individual infants stop making the error when they begin to self-locomote. Critically, one can take infants who do not yet self-locomote and who make the error and, by putting them in walkers, make them self-locomote 3 to 4 months earlier than they normally would. When one experimentally induces early experiences in self-locomotion, one also accelerates the development of successful search in the *A*-not-*B* task [4]. Why should experience in moving oneself about the world help one remember and discriminate the locations of objects in a hide-and-seek reaching task? The answer is because moving oneself about—over things, by things, into things, around things—presents new experiences, new patterns of spatiotemporal relations, that alter the infant's representation of objects, space, and self.

All in all, infants' experiences—the regularities the learning system encounters—change systematically as a function of development itself. Each developmental achievement on the part of the infant—hand-eye coordination, sitting, crawling, walking—opens the infant to whole new sets of multimodal regularities. Now here is the question that is critical for the *creation* of artificial life: Does the ordering of experiences matter in the final outcome? Could one just as well build an intelligent 2-year-old by starting with a baby that listened, looked, reached, and walked all together right from the beginning?

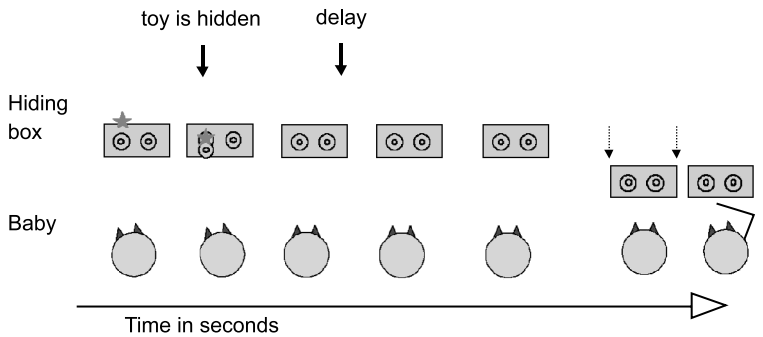

Figure 4. A schematic illustration of the course of events in the *A*-not-*B* task. After the delay, the hiding box is moved forward, allowing the infant to reach and search for the hidden toy.

Studies of comparative development make clear that the developmental ordering of sensory systems matters greatly. Different species show decidedly different orderings in the development of sensory systems, and these differences are related to their specific form of intelligence. The differential timing or heterochronicity is one way that evolution selects for different adaptive outcomes (see especially [53]). Experimental studies show that *reorderings*—changes in the normal development path—dramatically alter developmental outcomes. For example, opening kittens' eyes early disrupts olfactory development and the subsequent coordination of vision and olfaction [27, 38]. Similarly, disrupting the developmental order of audition and vision in owls disrupts spatial localization in both modalities [23]. One of the ingredients in building biological intelligence is ordering the training experiences in the right way.

Several attempts to model human learning [17, 35, 40] have shown that neural networks sometimes fail to learn the task when the entire data set is presented all at once, but succeed when the data are presented incrementally with an easy-to-difficult ordering. These demonstrations have been criticized by some as cheating. But to those of us who study how intelligence gets made in real live babies, they seem to have the right idea. Of course, in real development, this ordering of training experiences is not forced on the learner by some omnipresent teacher, but rather emerges as a consequence of development itself.

2.3 Lesson 3: Be Physical

Not all knowledge needs to be put into the head, into dedicated mechanisms, into representations. Some knowledge can be realized in the body, a fact dramatically illustrated by passive walkers. Knowledge of the alternating limb movement of bipedal locomotion—knowledge traditionally attributed to a central pattern generator—appears to reside in the dynamics of two coupled pendulums [30]. Some of our intelligence also appears to be in the interface between the body and the world. The phenomenon of change blindness is often conceptualized in this way. People do not remember the details of what is right before their eyes, because they do not need to remember what they can merely look at and see [32]. Similarly, Ballard and colleagues [2] have shown that in tasks in which people are asked to rearrange arrays of squares, they offload their short-term memory to the world (when they can). This offloading in the interface between body and world appears a pervasive aspect of human cognition and may be critical to the development of higher-level cognitive functions or in the binding of mental contents that are separated in time. We briefly present some new data that illustrate this point [45].

The experimental procedure derives from a task first used by Baldwin [1] and illustrated in Figure 5. The participating subjects are very young children, $1\frac{1}{2}$ to 2 years of age. The experimenter sits before a child at a table, and (a) presents the child with first one object to play with and then (b) with a second. Out of sight of the child, the two objects are then put into containers, and the two containers (c) are placed on the table. The experimenter looks into one container (d) and says, "I see a dax in here." The experimenter does not show the child the object in the container. Later the objects are retrieved from the containers (e) and the child is asked which one is "a dax." Notice that the name and the object were never jointly experienced. How then can the child join the object name to the right object? Baldwin showed that children as young as 24 months could do this, taking the name to refer to the *unseen* object that had been in the bucket at the same time the name was offered. How did children do this? How, if you were building an artificial device, would you construct a device that could do this, that could know the name applied to an object not physically present when the name was offered?

There are a number of solutions that one might try, including reasoning and remembering about which objects came out of which containers and about the likely intentions of speakers when they offer names. The evidence, however, indicates that young children solve this problem in a much simpler way, exploiting the link between objects and locations and space. What children do in this task is make use of a deep and foundationally important regularity in the world: A real object is perceptually distinguished from others based on its unique location; it must be in a different place from any other object. The key factor in the Baldwin task is that in the first part of the experimental

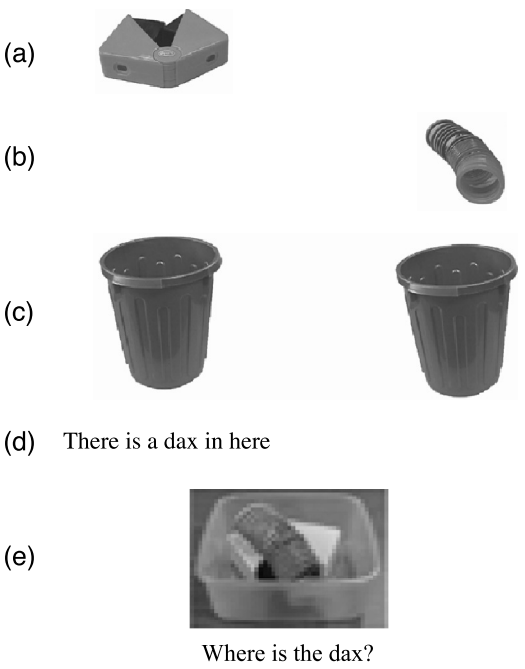

Figure 5. Events in the Baldwin task (see text for further clarification).

procedure, one object is presented on the right, the other on the left. The containers are also presented one on the right, one on the left, and the name is presented with attention directed (by the experimenter's looking into the bucket) to one location, for example, on the right. The child solves this task by linking the name to the object associated with that location. We know this is the case because we can modify the experiment in several crucial ways. For example, one does not need containers or hidden objects to get the result at all. One can merely present the target object on the right and have children attend to and play it with it there, then present the distracter object on the left and have children attend to and play with it there. Then, with all objects removed, with only an empty and uniform table surface in view, one can direct children's attention to the right and offer the name (dax) or to the left and offer the name. Children consistently and reliably link the name to the object that had been at that location.

Young children's solution to this task is simple, a trick in a sense, that makes very young children look smarter than they perhaps really are. But it is a trick that will work in many tasks. Linking objects to locations and then using attention to that location to link related events to that object provides an easy way to bind objects and predicates [2]. People routinely and apparently unconsciously gesture with one hand when speaking of one protagonist in a story and gesture with the other hand when speaking of a different protagonist. In this way, by hand gestures and direction of attention, they link separate events in a story to the same individual. American Sign Language formally uses space in this way in its system of pronouns. People also use space as a mnemonic, looking in the direction of a past event to help remember that event. One experimental task that shows this is the *Hollywood Squares* experiments of Richardson and Spivey [36]. People were presented at different times with four different videos, each from a distinct spatial location. Later, with no videos present, the subjects were asked about the content of those videos. Eye-tracking cameras recorded where people looked when answering these questions, and the results showed that they systematically looked in the direction where the relevant information had been previously presented.

This is all related to the idea of *deictic pointers* [2, 22] and is one strong example of how sensorimotor behaviors—where one looks, what one sees, where one acts—create coherence in our

cognition system, binding together related cognitive contents and keeping them separate from other distinct contents. In sum, one does not necessarily need lots of content-relevant knowledge or inferential systems to connect one idea to another. Instead, there is a cheaper way: by using the world and the body's pointers *to that world*.

2.4 Lesson 4: Explore

How can a learner *who does not know what there is to learn* manage to learn anyway? This is a more difficult question than it might first appear. The issue is whether one needs to prespecify the learning tasks and the learning goals: whether the agent or its designer has to know what needs to be learned in order to learn. Evidence from human development gets us out of this quandary by showing that babies can discover both the tasks to be learned and the solution to those tasks through exploration, or non-goal-directed action. In babies, spontaneous movement creates both tasks and opportunities for learning. One elegant demonstration concerns the study of reaching [11]. The week-by-week development of four babies was tracked over a 3-month period as they transitioned from not reaching to reaching. Four very different patterns of development were observed. Some babies in the non-reaching period hardly lifted their arms at all, but sat placidly watching the world. Other babies were more high-strung and active, flailing and flapping and always moving. These different babies had to learn to solve very different problems in order to learn to reach out and grasp an object. The flailer would have to learn to become less active, to lower his hands, to bring them into midline. The placid baby would have to learn to be more active, to raise her hands, to lift them up from their usual positions on her side. Each baby did learn, finding a solution that began with exploration of the movement space.

The course of learning for each baby appeared to be one of arousal, exploration, and the selection of solutions from that exploration space. In basic form, the developmental pattern is this: The presentation of an enticing toy is arousing and elicits all sorts of nonproductive actions, and very different individual actions in different babies. These actions are first, quite literally, all over the place with no clear coherence in form or direction. But by acting, by movements that explore the whole range of the movement space, each baby, in its own unique fashion, sooner or later makes contact with the toy—banging into or brushing against it or swiping it. These moments of contact select some movements in this space, carving out patterns that are then repeated with increasing frequency. Over weeks, the cycle repeats: arousal by the sight of some toy, action, and occasional contact. Over cycles, increasingly stable, more efficient, and more effective forms of reaching emerge. What is remarkable in the developmental patterns of the children is that each found a solution—and eventually converged to highly similar solutions—by following individually different developmental pathways. As they explored different movements, in their uncontrolled actions initiated by the arousing sight of the toy, they each discovered initially different patterns; each had a different developmental task to solve. The lesson for building intelligent agents is clear: A multimodal system that builds reentrant maps from time-locked correlations only needs to be set in motion, to move about broadly, even randomly, to learn and through such exploration to discover both tasks and solutions.

The power of movement as a means for exploration is also illustrated by an experimental procedure known as *infant conjugate reinforcement* [39]. Infants (as young as 3 months) are placed on their backs, and their ankles are attached by a ribbon to a mobile, which is suspended overhead. Infants, of course, through their own actions, discover this link. As the infants kick their feet, at first spontaneously, they activate the mobile. Within a few minutes they learn the contingency between their foot kicks and the jiggling of the mobile, which presents interesting sights and sounds. The mobile responds conjugately to the infants' actions: The more infants kick and the more vigorously they move, the more motion and sound they produce in the mobile. In this situation, infants increase their kicking to above the baseline spontaneous levels apparent when babies simply look at a non-moving mobile. Infants' behavior as they discover their control is one of initial exploration of a wide variety of actions and the selection of the optimal pattern to make the interesting events—the movement of the mobile—occur.

Although this is an experimental task, and not an everyday real-world one, it is a very appropriate model for real-world learning. The mobile provides the infant with many time-locked patterns of correlations. More importantly, infants themselves discover the relations through their own exploratory movement patterns. The infants themselves are moving contingently with the mobile; the faster and harder they kick, the more vigorously the mobile jiggles and sways. This is for infants a highly engaging task; they smile and laugh, and often become angry when the contingency is removed. Thus, the experimental procedure, *like the world,* provides complex, diverse, and never exactly repeating events, yet all perfectly time-locked with infants' own actions. And it is exploration, spontaneous non-task-related movement, that starts the process off. Without spontaneous movement, without exploration, there is nothing to learn from the mobile.

Young mammals, including children, spend a lot of time in behavior with no apparent goal. They move, they jiggle, they run around, they bounce things and throw them, and generally abuse them in ways that seem, to mature minds, to have no good use. However, this behavior, commonly called play, is essential to building inventive forms of intelligence that are open to new solutions.

2.5 Lesson 5: Be Social

Let us re-imagine the infant conjugate reinforcement paradigm. However, in this case instead of coupling the infant's leg by ribbon to a mobile, we couple the infant's face by mutual gaze to another face, the face of a mature partner. Many developmental researchers have observed mother-infant face-to-face interactions, and they report a pattern of activity and learning that looks very much like conjugate reinforcement, but with an added twist [9, 37, 43, 52]. Mothers' facial gestures and the sounds they make are tightly coupled to the babies' behavior. When babies look into their mother's eyes, mothers look back and smile and offer a sound with rising pitch. When babies smile, mothers smile. When babies coo, mothers coo. Babies' facial actions create interesting sights and sounds from mothers, just as their kicks create interesting sights and sounds from attached mobiles. And just as in the case of the ribbon-tethered mobiles, these contingencies create a context for arousal and exploration. In the initial moments as infants and mothers interact, infants' vocalizations and facial expressions become more active, broader, and more diverse. This exploration sets up the opportunity for learning time-locked correspondences between infants' facial actions and vocalizations and those of the mother, such that the infants' actions become transformed by the patterns they produce in others.

But crucially, the social partner in these adventures offers much more than a mobile, and this changes everything. Mature social partners do not just react conjugately to the infants' behavior; they build on it and provide scaffolding to support it and to transform it into conventionally shared patterns. For example, very early infant behavior shows a natural rhythmic pattern of intense excitement alternating with patterns of relative calm [9, 37, 43, 52]. Caregivers are thus able to create a conversation-like exchange by weaving their own behavior around the child's natural activity patterns. Initially, it appears that the caregiver alone is responsible for the structure of interaction. But babies' behaviors are both entrained by the mother's pattern and educated by the multimodal correspondences those interactions create. Incrementally and progressively, the babies become active contributors, affecting the mother by their own reactions to her behavior, and keeping up their own end of the conversation.

Imitation provides another example of the scaffolding mature partners provide to the developmental process. Although the evidence that babies reflexively imitate parental facial gestures at birth is controversial, other research does strongly suggest that infants learn to imitate parent vocalizations. Parents provide the structure for this learning by imitating their babies! That is, parents do not just respond to their infants' smiles and vocalizations; they imitate them. This sets up a cyclical pattern: vocalization by the infant, imitation by the parent, repeated vocalization by the infant, imitation by the parent, and so on. This creates opportunities for learning and fine-tuning the infant's facial and vocal gestures to match the adult model. In brief, the cycle works to strengthen and highlight certain patterns of production as parents naturally select those that they take to be meaningful [9, 29, 43].

Mature social partners also provide multimodal supports to help ground early language learning. When a parent introduces an object to a toddler and names it, the parent musters a whole array of sensorimotor supports to bring the child's attention to the object and to bind that object to the word [19–21]. Parents look at the object they are naming, they wave it so that the child will look at it, and they match the intonation patterns in which they present the name to their very actions in gesturing at or waving the object. In one study, Yoshida and Smith [54] observed that both English-speaking and Japanese-speaking parents routinely couple action and sound when talking to young children. For example, one parent demonstrated a toy tape measure to their child and when pulling the tape out said, "See, you pullllllll it," elongating the word pull to match the stopping and starting of the action of pulling. This same parent, when winding the tape back in, said, "Turn it round and round and round and round and round and round," with each "round" coinciding with the start of a new cycle of turning. By tying action and sound, parents ground language in the same multimodal learning processes that undergird all of cognition, and in so doing, they capture children's attention, rhythmically pulling it to the relevant linguistic and perceptual events, and tightly binding those events together.

Again the lesson for building intelligent agents is clear: Raise them in a social world, coupling their behavior and learning to agents who add structure and support to those coupled interactions.

2.6 Lesson 6: Learn a Language

Language appears to begin highly grounded in the perceptual here and now, to sensorimotor and social processes that are not specific to language but rather quite open learning systems, capable of discovering and solving an infinite variety of tasks. But language is — without a doubt — a very special form of regularity in the world, and one that profoundly changes the learner.

First, language is an in-the-world regularity that is a *shared* communicative system [18]. Its shared aspect means that it is very stable, continually constrained by the many local communicative acts of which is composed. As an emergent product made out of many individual communicative encounters, the structure of natural languages may be nearly perfectly adapted to the learning community. At any rate, in the lives of humans, language is as pervasive, as ubiquitous in its role in intelligence, as is gravity.

Second, language is special because it *is* a symbol system. At the level of individual words (morphemes really), the relation between events in the world and the linguistic forms that refer to them is mainly arbitrary. That is, there is no intrinsic similarity between the sound of most words and their referents: the form of the word *dog* gives us no hints about the kinds of thing to which it refers. And nothing in the similarity of the forms of *dig* and *dog* conveys a similarity in meaning. It is interesting to ask *why* language is this way. One might expect that a multimodal, grounded, sensorimotor sort of learning would favor a more iconic, pantomime-like language in which symbols were similar to referents. But language is decidedly not like this. Moreover, the evidence suggests that although children readily learn mappings supported by multimodal iconicity, they fail if there is *too much* iconicity between the symbol and the signified.

One intriguing demonstration of this comes from the research of DeLoache [12], which is directed not to language learning, but to children's use of scale models. DeLoache's experimental task is a hiding game, and the children are 2-year-olds. On each trial, a toy is hidden in a real life-size room, say, under a couch. The child's task is to find the toy, and on every trial the experimenter tells the child exactly where the toy is, using a model of some kind. This model might be a blueprint, a drawing of the room, a photograph, a simple scale model, a richly detailed and exact scale model, or a life-size model. Here is the very robust but counterintuitive result: Young children fail in this task whenever the model is too similar to the real room. For example, they are much more likely to succeed when the solution is shown in a picture than in a scale model and much more likely to succeed when the scale model is a simplified version of the real room than an accurate representation. Why mustn't a symbol be too lifelike, too much like the real world? One possibility is that children must learn what a symbol is, and to learn what a symbol is, there must be some

properties that are common to the set of symbols, for example, the properties that distinguish pictures from real objects or spoken words from other sounds.

The fact that all the world's languages are symbol systems, the fact that too much similarity between a symbol and the signified disrupts the learning of a mapping between them, suggests that arbitrary symbols confer some unique and valuable computational power. That power might lie in the property of *orthogonality*. For the most part, individual words pick out or point to unique categories. At the very least, this is true in the lexicons of 2- to 3-year-olds [42]. We also know that young children act as if it were true, a phenomenon sometimes referred to as the mutual exclusivity constraint [8, 28]. More specifically, children act as if each object in the world received one and only one name. For example, shown two novel objects and told the name of one (e.g., "This is a dax"), children will assume that any new name (e.g., "wug") refers to the second previously unnamed object. The arbitrariness and mutual exclusivity of linguistic labels may be computationally powerful because they pull the overlapping regularities that create perceptual categories apart, as illustrated in Figure 6. There is evidence to support this idea that orthogonality is computationally powerful, enabling children to form second-order, rule-like generalizations. To explain this developmentally powerful aspect of language learning, we must first provide some background on children's word learning.

Children comprehend their first word at around 10 months; they produce their first word at around 12 months. Their initial progress in language learning is surely built on multimodal clusters and categories emergent in the infant's interactions in the world. Nonetheless, progress at first is

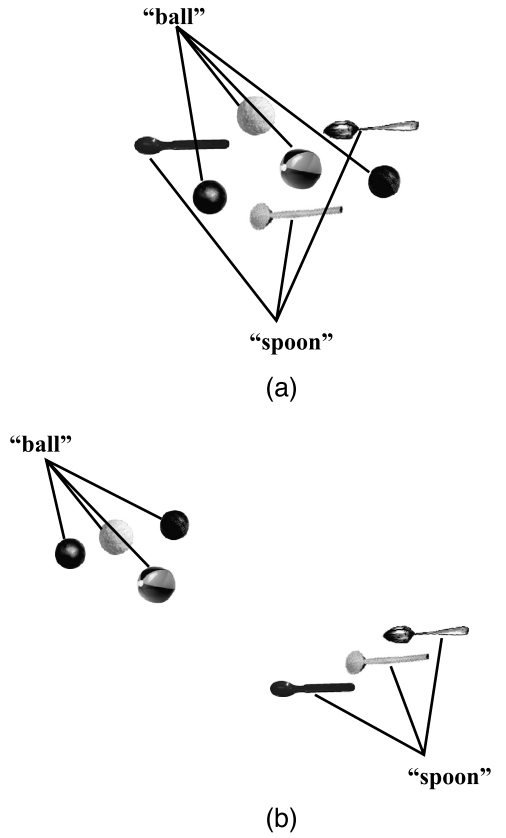

Figure 6. Illustration of Colunga's [10] proposal of how orthogonal labels may pull similarities apart. (a) Word forms become associated with members and features of object categories. (b) The orthogonality of the words leads to the divergence of initially similar conceptual representations.

tentative, slow, and fragile. For the 6 months or longer after the first word, children acquire subsequent words very slowly, and often seem to lose previously acquired ones. Moreover, they seem to need to hear each individual word in many contexts before they apprehend its range. Then, between 18 and 20 months, most children become very rapid word learners, adding new words to their vocabularies at the staggering rates of 4 to 9 a day. During this time, they seem to need only to hear a word used to label a single object to know the whole class of things to which the word refers [44]. This one-instance to whole-category learning is especially remarkable in that different kinds of categories are organized in different ways. For example, animate categories are organized by many different kinds of similarities across many modalities; artifact categories are organized by shape, and substance categories by material.

The evidence from both experimental studies and computational models indicates that children learn these regularities as they slowly learn their first words and that this learning then *creates* their ability to learn words in one trial. The nature of this learning can be characterized by four steps, illustrated in Figure 7 [41, 46]. The figure illustrates just one of the regularities that children learn: that artifact categories are organized by shape. Step 1 in the learning process is the mapping of names to objects—the name "ball" to a particular ball and the name "cup" to a particular cup, for example. This is done multiple times for each name as the child encounters multiple examples. And importantly, in the early lexicon, solid, rigidly shaped things are in categories typically well organized by similarity in shape [42]. This learning of individual names sets up step 2—first-order generalizations about the structure of individual categories, that is, the knowledge that balls are round and cups are cup-shaped. The first-order generalization should enable the learner to recognize novel balls and cups.

Another higher-order generalization is also possible. Because most of the solid and rigid things that children learn about are named by their shape, children may also learn the second-order generalization that names for artifacts (solid, rigid things) in general span categories of similar-shaped things. As illustrated in step 3 of the figure, this second-order generalization requires generalizations over specific names and specific category structures. But making this higher-order generalization should enable the child to extend any artifact name, even one encountered for the first time, to new instances by shape. At this point, the child behaves as if it has an abstract and variablized rule: For any artefact, whatever its individual properties or individual shape, form a category by a shape. Step 4 illustrates the potential developmental consequence of this higher-order generalization—attention to the right property, shape—for learning new names for artifacts. The plausibility of this account has been demonstrated in experimental studies that effectively accelerate the vocabulary acquisition function by teaching children the relevant correlations and in simulation studies with neural nets [41, 46].

How special is language's role in enabling the formation of second-order generalizations? Perhaps very special indeed. Recent simulation studies by Colunga [10] suggest that the arbitrariness and orthogonality of the linguistic labels may be critical. Neural networks that readily form second-order generalizations and yield accelerating rates of vocabulary acquisition do not do this if the labels, the words, are not orthogonal. We strongly suspect that even if orthogonality does not prove in the limit to be necessary, it will prove to be strongly beneficial to the formation of second-order generalizations. This work is a beginning hint at an answer to what we take to be a deeply important question for those of us who wish to understand intelligent systems: Why in a so profoundly multimodal sensorimotor agent such as ourselves is language an arbitrary symbol system? What computational problems are solved by language taking this form?

The advantages conferred by arbitrary symbols go well beyond those hinted at here. More familiar are the properties of symbol *systems*, capacities that result from the possibility of combining symbols. For natural languages, this is the domain of grammar. All known natural languages have two fundamental properties of symbol systems.

First, they are at least approximately compositional. That is, in the domain of grammar, unlike the domain of individual morphemes, language is anything but arbitrary. Compositionality permits hearers to comprehend unfamiliar combinations of morphemes and speakers to produce combina-

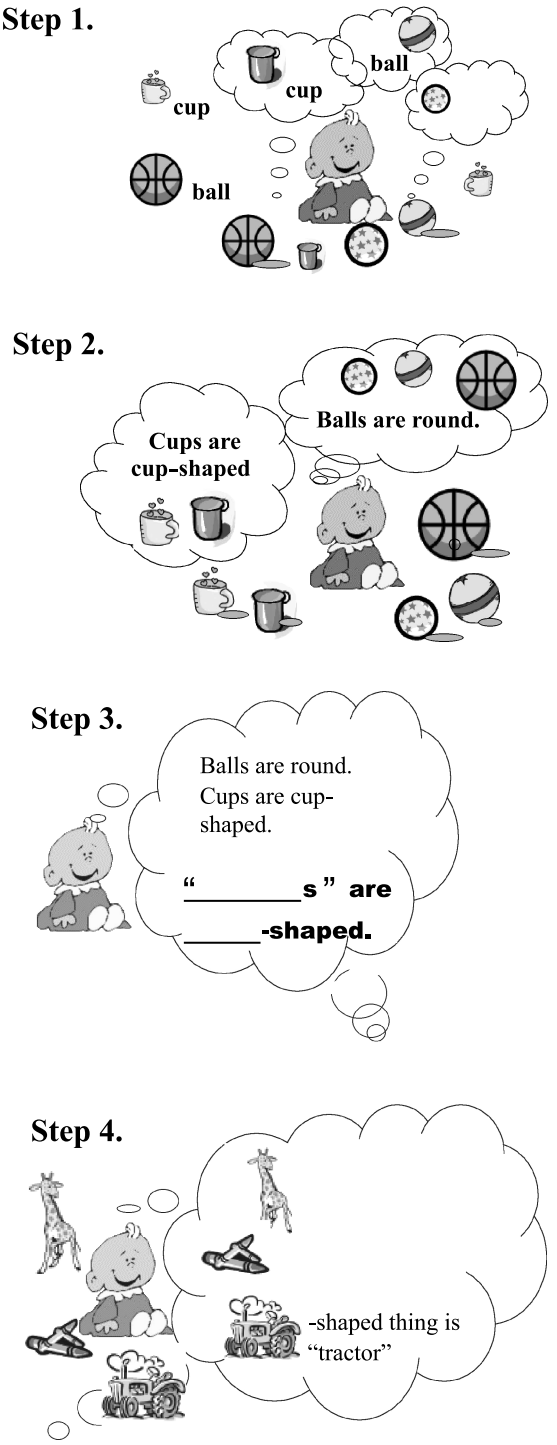

Figure 7. Four steps in the development of one-trial word learning. Step 1: mapping of names to objects. Step 2: first-order generalizations about the structure of individual categories. Step 3: second-order generalization. Step 4: attention to the right property in learning new names.

tions of morphemes they've never produced or heard before. An English speaker who knows what a *dax* is automatically knows that *daxes* refers to more than one of them.

Second, words as symbols permit structured representations, in particular, those with embedding. Embedding is possible because symbols representing relations between symbols can themselves play the role of symbols representing objects. So we can say things like *John thinks Mary doubts he likes her* and *the woman who teaches the class I like.*

It may be the orthogonal nature of linguistic representations, deriving ultimately from the arbitrary nature of the form-meaning relationship at the level of morphemes, that is behind these properties of language as well. If the representations for the words in a sentence overlapped significantly, it would be impossible to keep them separate in composing the meanings of the words. Orthogonal representations permit several separate items to be maintained simultaneously in short-term memory without significant interference. This does not deny the rich, distributed representations for the concepts behind these words; it simply brings out the value of orthogonal *pointers* to those representations. These pointers can be manipulated (composed, associated in structures) without making direct reference to their meanings or their pronunciations.

But this power goes beyond the grammar of natural languages. This same potential for composition and for structured representations holds for other symbolic processing generally and seems to characterize human activities such as explicit planning and mathematics. It has long been suggested that the way into symbolic processing is through language [56], and though this idea remains controversial, we believe it is worth taking seriously. First, because sentence structure maps onto event structure, language could teach children about how to attend to event structure in the same way that it apparently teaches them to attend to particular dimensions of objects. Second, the orthogonal symbols that allow language to be compositional and structured, once learned, could provide the basis for other symbol systems, such as the one that is behind algebra.

Developing in a linguistic world makes children smarter in at least three kinds of ways. First, and most obviously, by learning a language, children gain more direct access to the knowledge that others have. Children can be instructed, and when they are unsure of something, they can ask questions and can eventually search for the information in written form. While knowledge in this explicit verbal form may not have the richness of knowledge that results from direct experience, it can supplement the experience-based knowledge, especially in areas where children have no possibility of direct experience.

Second, in learning a language, children are presented with an explicit categorization of the objects, attributes, and relations in the world. Each morpheme in a natural language represents a generalization over a range of sensory, motoric, and cognitive experiences, and by labeling the range, morphemes function as a form of supervised category learning that is unavailable to other organisms. Thus, one result of learning a language is an ontology. Not only does this permit children to notice regularities they might miss otherwise (for example, the relevance of shape for artifacts or motion for animates), but because the ontology is shared by the community of speakers of the language, it guarantees a degree of commonality in the way the members of the community will respond to the world.

Third, and as we suggested here, learning a language may be the key to becoming symbolic and by its very nature may change the computational power of the learner. Each word associates a distributed phonological pattern and a distributed conceptual pattern in what is apparently a local, or at least orthogonal, fashion. It may be the largely arbitrary nature of this association that facilitates the learning of local lexical representations; because similarity in word forms does not entail similarity in the corresponding meanings, and vice versa, mediating representations that do not overlap are the most efficient alternative. Whatever the reason, research on lexical access in language production [14, 26] points, first, to the psychological reality of a distinct lexical level of representation and, second, to the fundamentally orthogonal and competitive nature of these representations. The advantage of these local representations is that complex reasoning can be carried out on them directly: They can be associated with one another and even arranged in hierarchical structures, representing symbolically what could not be achieved with the distributed overlapping representa-

tions of component concepts. Thus the power of symbolic reasoning—planning, logic, and mathematics—may derive ultimately from words in their function as pointers to concepts.

3 Conclusion

Artificial life attempts to model living biological systems through complex algorithms. We have suggested in this article that developmental psychology offers usable lessons for creating the intelligence that lives in the real world, is connected to it, and knows about that world. Babies begin with a body richly endowed with multiple sensory and action systems. But a richly endowed body that is simply thrown into a complex world, even with the benefits of some pre-programming and hardwiring by its designers, would fail to meet the standard of even a 3-year-old unless it were tuned to the detailed statistics of that world. We have argued that embodied intelligence *develops*. In an (embodied) human child, intelligence emerges as the child explores the world, using its sophisticated statistical learning abilities to pick up on the subtle regularities around it. Because the child starts small, because its intelligence builds on the progress it has already made, because development brings the child to different regularities in the world, because those regularities include couplings between the child and smart social partners, and because the world includes a symbol system, natural language, the child achieves an intelligence beyond that of any other animal, let alone any current artificial device. The lesson from babies is: intelligence isn't just embodied; it *becomes* embodied.

Acknowledgments

The preparation of this manuscript and much of the work reported in it were supported by NIH-NIMH R01MH60200.

References

1. Baldwin, D. (1993). Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental Psychology*, *29*(5), 832–843.
2. Ballard, D., Hayhoe, M., Pook, P., & Rao, R. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, *20*, 723–767.
3. Barsalou, L. W. (in press). Abstraction as a dynamic construal in perceptual symbol systems. In L. Gershkoff-Stowe & Rakison, David (Eds.), *Building object categories in developmental time*. Hillsdale, NJ: Erlbaum.
4. Bertenthal, B., Campos, J., & Barrett, K. (1984). Self-produced motion: An organizer of emotional, cognitive, and social development in infancy. In R. Emde & R. Harmon (Eds.), *Continuities and discontinuities* (pp. 175–210). New York: Plenum Press.
5. Breazeal, C. (2002). *Designing sociable robots*. Cambridge, MA: MIT Press.
6. Brooks, R., Breazeal, C., Marjanovic, M., Scassellati, B., & Williamson, M. (1998). The cog project: Building a humanoid robot. In C. Nehaniv (Ed.), *Computation for metaphors, analogy and agents*. Springer-Verlag.
7. Bushnell, E. (1994). A dual processing approach to cross-modal matching: Implications for development. In D. Lewkowicz & R. Lickliter (Eds.), *The development of intersensory perception* (pp. 19–38). Mahwah, NJ: Erlbaum.
8. Clark, E. (1987). The principle of contrast: A constraint on language acquisition. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 1–33). Hillsdale, NJ: Erlbaum.
9. Cohn, J. F., & Tronick, E. Z. (1988). Mother-infant face to face interaction: Influence is bi-directional and unrelated to periodic cycles in either partner's behavior. *Developmental Psychology*, *24*, 386–392.
10. Colunga, E. (2003). Local vs. distributed representations: Implications for language learning. In preparation.
11. Corbetta, D., & Thelen, E. (1996). The developmental origins of bimanual coordination. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 502–522.
12. DeLoache, J. (2002). The symbol-mindedness of young children. In W. Hartup & R. Weinberg (Eds.),

Child psychology in retrospect and prospect: In celebration of the 75th anniversary of the Institute of Child Development (pp. 73–101). Mahwah, NJ: Erlbaum.

13. Diamond, A. (1990). Developmental time course in human infants and infant monkeys and the neural bases of inhibitory control in reaching. In A. Diamond (Ed.), *The development and neural bases of higher cognitive functions* (pp. 637–676). New York: New York Academy of Sciences Press.
14. Dell, G. S, Juliano, C., & Govindjee, A. (1993). Structure and content in language production: A theory of frame constraints in phonological speech errors. *Cognitive Science*, *17*, 149–195.
15. Edelman, G. (1987). *Neural Darwinism*. New York: Basic Books.
16. Ellis, R., & Tucker, M. (2000). Micro-affordance: The potentiation of components of action by seen objects. *British Journal of Psychology*, *91*(4), 451–471.
17. Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*, 71–99.
18. Freyd, J. (1983). Shareability: The social psychology of epistemology. *Cognitive Science*, *7*(3), 191–210.
19. Gogate, L., & Bahrick, L. (2001). Intersensory redundancy and 7-month-old infants' memory for arbitrary syllable-object relations. *Infancy*, *2*(2), 219–231.
20. Gogate, L., & Walker-Andrews, A. (2001). More on developmental dynamics in lexical learning. *Developmental Science*, *4*(1), 31–37.
21. Gogate, L., Walker-Andrews, A., & Bahrick, L. (2001). The intersensory origins of word comprehension: An ecological-dynamic systems view. *Developmental Science*, *4*(1), 1–18.
22. Hurford, J. (in press). The neural basis of predicate-argument structure. *Behavioral and Brain Sciences*.
23. Knudsen, E. (2003). Instructed learning in the auditory localization pathway of the barn owl. *Nature*, *417*(6886), 322–328.
24. Lakoff, G. (1994). What is a conceptual system? In W. F. Overton & D. S. Palermo (Eds.), *The nature and ontogenesis of meaning. The Jean Piaget symposium series* (pp. 41–90). Hillsdale, NJ: Erlbaum.
25. Landau, B., & Gleitman, L. (1985). *Language and experience*. Cambridge, MA: Harvard University Press.
26. Levelt, W. J. M. (2001). Spoken word production: A theory of lexical access. *Proceedings of the National Academy of Sciences*, *98*, 13,464–13,471.
27. Lickliter, E. (1993). Timing and the development of perinatal perceptual organization. In G. Turkewitz & D. Devenney (Eds.), *Developmental time and timing* (pp. 105–123). Hillsdale, NJ: Erlbaum.
28. Markman, E., & Wachtel, G. (1988). Children's use of mutual exclusivity to constrain the meaning of words. *Cognitive Psychology*, *20*(2), 121–157.
29. Masur, E., & Rodemaker, J. (1999). Mothers' and infants' spontaneous vocal, verbal, and action imitation during the second year. *Merrill-Palmer Quarterly*, *45*(3), 392–412.
30. McGeer, T. (1990). Passive dynamic walking. *International Journal of Robotics Research*, *9*(2), 62–82.
31. Mendelson, M. J., & Haith, M. M. (1976). The relation between audition and vision in the human newborn. *Monographs of the Society for Research in Child Development*, *41*(4); Serial No. 167.
32. O'Regan, J. K., & Noë, A. (in press). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, *24*, 939–973.
33. Pfeifer, R., & Scheier, C. (1999). *Understanding intelligence*. Cambridge, MA: MIT Press.
34. Piaget, J. (1963). *The origins of intelligence in children*. New York: Norton.
35. Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. *Cognition*, *38*, 1–60.
36. Richardson, D., & Spivey, M. (2000). Representation, space, and Hollywood Squares: Looking at things that aren't there anymore. *Cognition*, *76*, 269–295.
37. Rogoff, B. (1990). *Apprenticeship in thinking: Cognitive development in social context*. Oxford, UK: Oxford University Press.
38. Rosenblatt, J. S., Turkewitz, G., & Schneirla, T. C. (1969). Development of home orientation in newborn kittens. *Transactions of the New York Academy of Sciences*, *31*, 231–250.

39. Rovee-Collier, C., & Hayne, H. (1987). Reactivation of infant memory: Implications for cognitive development. In H. Reese (Ed.), *Advances in child development and behavior*, *Vol. 20*, (pp. 185–238). San Diego, CA: Academic Press.

40. Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tense of English verbs. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition*, *Volume 2*, Cambridge, MA: MIT Press.

41. Samuelson, L. (2002). Statistical regularities in vocabulary guide language acquisition in connectionist models and 15–20-month olds. *Developmental Psychology*, *38*, 1011–1037.

42. Samuelson, L. K., & Smith, L. B. (1999). Early noun vocabularies: Do ontology, category structure and syntax correspond?. *Cognition*, *73*(1), 1–33.

43. Schaffer, H. R. (1996). *Social development*. Oxford, UK: Blackwell.

44. Smith, L. B. (2000). How to learn words: An associative crane. In R. G. K. Hirrsh-Pasek (Ed.), *Breaking the word learning barrier* (pp. 51–80). Oxford, UK: Oxford University Press.

45. Smith, L. B. (2003). *How space binds words and referents*. In preparation.

46. Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*(1), 13–19.

47. Smith, L. B., Quittner, A. L., Osberger, M. J., & Miyamoto, R. (1998). Audition and visual attention: The developmental trajectory in deaf and hearing populations. *Developmental Psychology*, *34*(5), 840–850.

48. Smith, L. B., Thelen, E., Titzer, R., & McLin, D. (1999). Knowing in the context of acting: The task dynamics of the A-not-B error. *Psychological Review*, *106*(2), 235–260.

49. Spelke, E. S. (1979). Perceiving bi-modally specified events in infancy. *Developmental Psychology*, *15*, 626–636.

50. Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

51. Thelen, E., Schoener, G., Scheier, C., & Smith, L. B. (2001). The dynamics of embodiment: A field theory of infant perseverative reaching. *Behavioral & Brain Sciences*, *24*(1), 1–86.

52. Trevarthen, C. (1988). Infants trying to talk. In R. Söderbergh (Ed.), *Children's creative communication*. Lund, Sweden: Lund University Press.

53. Turkewitz, G., & Kenny, P. A. (1985). The role of developmental limitations of sensory input on sensory/perceptual organization. *Journal of Developmental and Behavioral Pediatrics*, *6*, 302–306.

54. Yoshida, H., & Smith, L. B. (2003). Sound symbolism and early word learning in two languages. Submitted to Annual Conference of the Cognitive Science Society.

55. Titzer, R., Thelen, E., & Smith, L. B. (2003). *Learning about transparency*. Unpublished manuscript.

56. Vygotsky, L. S. (1962). *Thought and language*. New York: MIT Press and Wiley.

57. Wertheimer, M. (1961). Psychomotor coordination of auditory-visual space at birth. *Science*, *134*, 1692.