TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 3 (2011) 686–706

ISSN: 1756-8757 print / 1756-8765 online
DOI: 10.1111/j.1756-8765.2011.01158.x

Embodied Spatial Cognition

J. Gregory Trafton, Anthony M. Harrison

Naval Research Laboratory, Washington, D.C.

Received 7 December 2009; received in revised form 15 April 2011; accepted 3 March 2011

Abstract

We present a spatial system called Specialized Egocentrically Coordinated Spaces embedded in an embodied cognitive architecture (ACT-R Embodied). We show how the spatial system works by modeling two different developmental findings: gaze-following and Level 1 perspective taking. The gaze-following model is based on an experiment by Corkum and Moore (1998), whereas the Level 1 visual perspective-taking model is based on an experiment by Moll and Tomasello (2006). The models run on an embodied robotic system.

Keywords: Spatial cognition; Gaze-following; Level 1 visual perspective taking; Cognitive robotics

1. Introduction

How do children develop spatial competence? Spelke and her colleagues have suggested that children have four sets of core knowledge: object representation, agents and their actions, number, and space (Dehaene, Izard, Pica, & Spelke, 2006; Spelke, 2003; Spelke & Kinzler, 2007). They suggest that even very young children have knowledge about the geometry of the environment, specifically the distance, angle, and relations among objects in a surrounding layout. If young children have so much core knowledge about space and geometry, an important issue is determining how their sense of space develops and how they integrate that different information into productive spatial cognition. Newcombe and Huttenlocher (2000) suggest that the development of spatial cognition occurs as different types of spatial information come into conflict. This conflict allows children and adults to learn which spatial cues to attend to in different situations, and then to use that knowledge to solve spatial problems. They also suggest that these changes would be represented as ‘‘reweighting’’ in a computational model (Newcombe & Huttenlocher, 2000, p. 49).

Correspondence should be sent to J. Gregory Trafton, Naval Research Laboratory, Code 5515, Washington, DC 20375. E-mail: greg.trafton@nrl.navy.mil

Gaze-following is one of the earliest forms of spatial cognition and is a key component of joint visual attention, or looking at the same object as another person (Butterworth & Jarrett, 1991; Scaife & Bruner, 1975). Previous researchers have suggested that joint visual attention is strongly related to the ability to infer others' mental states (Baron-Cohen, 1995).

A typical experiment (Butterworth & Jarrett, 1991; Corkum & Moore, 1998) in gaze-following has an infant come in and sit on a parent's lap. The parent's eyes are typically blindfolded. The experimenter sits directly across from the infant and calls the infant's name to get his or her attention. After the infant looks at the experimenter, the experimenter gazes at a toy in the room; if the infant also gazes at the same toy, the infant is said to follow the experimenter's gaze. If, instead, the infant looks elsewhere (e.g., at a distracter toy), the infant does not follow gaze on that trial (see Fig. 1a.)

Many researchers emphasize the social aspects of gaze-following (Baron-Cohen, 1995; Scaife & Bruner, 1975), but there are also many researchers who see it as one of the first examples of spatial cognition emerging in the infant (Butterworth & Jarrett, 1991; Corkum & Moore, 1995). Indeed, some models of gaze-following (e.g., Triesch, Teuscher, Deak, & Carlson, 2006) have been strongly critiqued for not modeling spatial cognition at an appropriate developmental level (Moore, 2006). For simple gaze-following, the infant must be able to view the experimenter's face and determine which direction the head or eyes (Brooks & Meltzoff, 2002) are oriented. The infant must also follow that direction in 3D space until a relevant or interesting object is found. Both determining face orientation and the process of mentally drawing a line between the face and an object in 3D space are clear examples of spatial cognition (Butterworth & Jarrett, 1991; Shepard & Metzler, 1971).

As infants progress beyond gaze-following, they start being able to take another's visual perspective. An initial form of this is when children are able to understand that the content of what they see may differ from the content of what another sees in the same situation, called Level 1 visual perspective taking (Flavell, 1999). One paradigm uses an occluder blocking the line of sight for the child or the experimenter to a toy (Moll & Tomasello, 2006). In this situation, the experimenter looks for a toy that is behind the occluder and asks the child if he or she knows where it is. The child succeeds at this task when the child gives the hidden toy to the experimenter; Fig. 1b shows an example of this paradigm.

Again, there are both social and spatial components of this task. In this Level 1 perspective-taking task, not only must orientation and location of the experimenter's head be used to determine what can be seen by the experimenter, but a spatial representation of where in space an object is relative to the child, including range information (Pick & Rieser, 1982), is

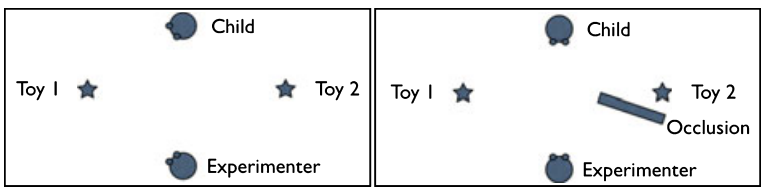

Fig. 1. Experimental setups for (A) Corkum and Moore (1998) and (B) Moll and Tomasello (2006).

also needed. Specifically, the child must determine whether the occluder is in between the toy and the experimenter—not too far away and not too close. This spatial information is critical for solving this class of problems.

In addition to the spatial representations needed for both the pure gaze-following and the Level 1 perspective-taking experiments, learning mechanisms and integrative processes are also needed. These paradigms are an excellent testbed for the intersection between learning, development, and spatial cognition.

In the remainder of this paper, we present an embodied approach to spatial cognition and how our models perform both gaze-following and Level 1 perspective-taking tasks. There are several reasons why we believe that spatial cognition models should be embodied. First, spatial cognition can be an enabler for other cognitive capabilities when embodiment is taken seriously (e.g., Wilson, 2002). Picking up an object, remembering where the keys are, tracking people as they move around, searching for a light switch, and so on can be (and have been) studied without spatial cognition. However, each of these tasks has a fundamental spatial component associated with it. Running models on an embodied platform not only forces the integration of different cognitive capacities (e.g., manipulation, memory, object tracking, visual search, time prediction) but also highlights where and which spatial processing mechanisms are critical.

Second, while simulations or virtual reality can provide a great testbed for some aspects of cognition, one of the reasons they are so useful is also a huge pitfall: they simplify the environment so much that models that interact with them can be fundamentally wrong. For example, a spatial model that interacts in a 2D world may completely break down in the 3D world people actually live in. Or a simulation that does not provide information on moving objects (or provides perfect information) can cause a cognitive model to mis- or underrepresent the visual and spatial cues that people actually use. Running models on embodied platforms provides a much stronger test of the theory than running the model solely in simulation. As Brooks and Matairc eloquently put it, ‘‘Simulations are doomed to succeed’’ (Brooks & Mataric, 1993, p. 209).

2. Architecture description

Because visual/spatial problem solving involves much more than simple manipulations of visual/spatial information, we have integrated our spatial theory with a more general account of problem-solving processes. Specifically, we have implemented our spatial module within ACT-R.

ACT-R is a hybrid symbolic/subsymbolic production-based system (Anderson, 2007). ACT-R consists of a number of modules and buffers, integrated with a central pattern matcher. Modules contain relatively specific cognitive faculties typically associated with a neurological network. Each module provides one or more buffers, which serve as interfaces between the module and ACT-R’s central production system. At any point in time, there may be at most one item in any individual buffer; thus, the module’s job is to decide what and when to put a symbolic object into a buffer. The pattern matcher uses the contents of

the buffer to match specific productions. Every 50 ms, a production rule can fire based on what is in each of the buffers. If there is more than one rule that can match the contents of the buffers, the rule that has the highest expected utility actually fires, makes a request of the module(s), which then carries out the associated actions.

ACT-R supports the concept of purely bottom-up processing. Bottom-up or reactive processing occurs when there is no goal-directed processing. In contrast, top-down or goal-directed processing occurs when the goal buffer (intentional module) is directing the processing.

ACT-R interfaces with the outside world through the visual (seeing), aural (hearing), motor (physical actions), and vocal (speaking) modules. Other current modules include the intentional (goal-directed), imaginal (problem state), temporal (sense of time), and declarative modules (fact memory). Each module deposits the results of its internal processing into the buffer, where it can be accessed by the central pattern matcher. For example, if the retrieval buffer is queried for the most active element given the current context and goal, the declarative module searches its memory for the declarative memory element that has the highest activation and puts it in the retrieval buffer. Issues like how long it takes to retrieve a specific element or whether or not there is a memory failure is handled by a set of underlying computational mechanisms in the declarative module (Anderson, Bothell, Lebiere, & Matessa, 1998).

Currently, ACT-R's spatial competencies are limited to the visual and aural modules. These modules provide some very simple spatial information (e.g., left, right, top, down, near). There is substantial neurological and psychological evidence that visual and spatial systems are distinct but interconnected to each other (Klatzky, 1998; McNamara & Shelton, 2003), and there have been several proposed systems to develop spatial competence in ACT-R (summarized in Gunzelmann & Lyon, 2007). As both gaze-following and Level 1 perspective taking require greater spatial competence than traditional ACT-R affords, we have added a set of spatial modules to ACT-R (the manipulative and configural modules). We have additionally augmented ACT-R by allowing it to perceive and act upon the physical world by interfacing the existing perceptual modules with robotic sensors and effectors; we call our system ACT-R/E (the ''E'' is for Embodied). We did not modify other parts of the architecture itself. Below we discuss the changes we made to the architecture. Fig. 2 shows a schematic of ACT-R/E.

2.1. Specialized Egocentrically Coordinated Spaces (SECS)

SECS[1] (pronounced SEEKS; Harrison & Schunn, 2002; Harrison, 2007) is a neurologically inspired, computational theory of the visual/spatial representational and computational abilities of the human mind. It integrates neuroscientific understanding of how the human brain represents visual/spatial information (Previc, 1998) with an analysis of the computational requirements of common visual/spatial tasks—in other words, it was built to be consistent with neuroscience but also to be computationally sufficient to do the tasks being described. SECS specifically posit that there are three different visual/spatial representational systems. The three representations make use of different (but interconnected)

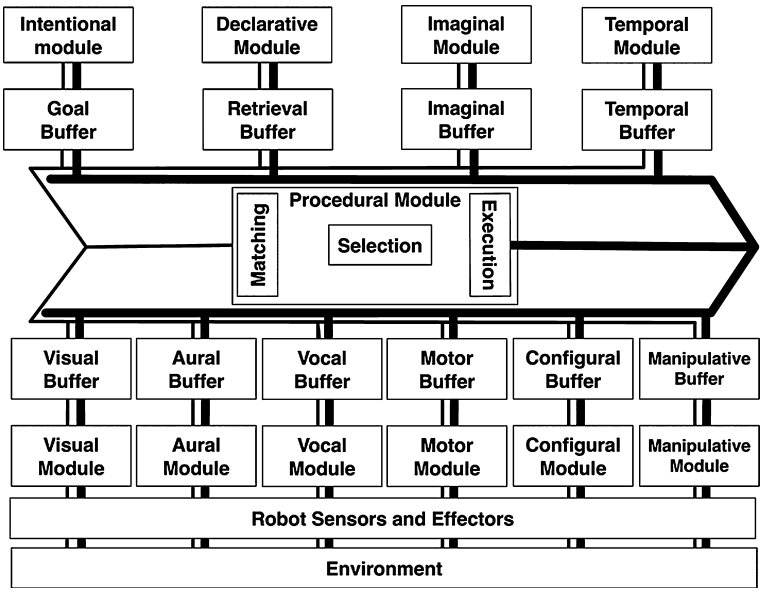

Fig. 2. Schematic of ACT-R/E.

neurological networks, tend to get used for different kinds of basic perceptual/motor tasks, have fundamentally different ways of representing space, and have different computational strengths and weaknesses. Note that the spatial systems are multimodal in that they integrate spatial information coming from the visual, auditory, haptic, proprioceptive, and vestibular systems, though we use only visual information in this project.

In SECS, the choice of which representation is used will be influenced by input: things in flat displays will tend to start out as visual; proximal objects will tend to start out as manipulative; and distal ones will tend to start out as configural. The reason for this is because of the strong ties between the visual and the spatial representations. SECS does, however, propose that people can translate between the three spatial representations. For example, a person can take arrangements of distant objects presumably represented configurally and translate it into a miniature manipulative 3D model or a flat map-like visual image. Additionally, people can create a visual, manipulative, or configural representation of any spatial object. Thus, a person can create a manipulative representation of a spatial object from across the room to determine its orientation. These representational changes are effortful and occur through requests made to the modules.

In general, an object starts off in a spatial representation based on its input, then the person can change its spatial representation based on task demands. For example, if a hammer were presented on a table for identification, it would be represented with a visual representation. If, however, a hammer were presented on a computer screen and the task was to decide which hand to pick it up with, the hammer would need to be translated to a manipulative representation. This translation process does not impact any module that is not involved in the translation process itself (e.g., visual, configural, or manipulative). So if additional

information is needed about the hammer, a declarative retrieval can be made, but that information is not needed while the translation process itself is occurring.

The implementation of SECS introduces new functionality to ACT-R. At the most basic level, it introduces two sets of spatial representations for the manipulative and configural systems and the transformations that can be applied to them (discussed below). The multimodal nature of SECS requires it to closely integrate with ACT-R's motor and visual systems. For example, when an ACT-R model visually attends to an object, the resulting symbolic representation becomes the link that ties the percept to the various spatial representations. As the model moves around the world (using the motor system), both the visual and spatial systems are dynamically updated, which allows path integration, manipulation, etc.

2.1.1. Visual

The visual module is used to provide a model with information about what can be seen in the current environment. The visual system is used for object identification, represents information primarily around the region that the eyes are attending to, and represents information in approximate shape terms and approximate size and location. Historically, this buffer has been called the ''What'' (ventral) visual pathway (Ungerleider & Mishkin, 1982). Its representation of the world is primarily 2D, with objects occupying space in the fronto-parallel plane (e.g., on a computer screen or chart on the wall in front of you). That is, there are approximate above/below and left/right relationships, but no strong distance nor exact orientation information. The visual representation is associated with the visual temporal cortex and fusiform gyrus (Anderson, Qin, Jung, & Carter, 2007).

We modified the original visual module to accept input from a video camera. The visual module allows access to object identification through fiducial (Kato, Billinghurst, Poupyrev, Imamoto, & Tachibana, 2000) and face (Fransen, Hebst, Harrison, & Trafton, 2009) trackers. Generally, these two systems are functional systems that provide object identification in a real-world environment. The fiducial tracker, for example, requires a small amount of training for individual, arbitrary symbols printed on an object. After training, that symbol can then be visually recognized as a labeled object (e.g., a toy). The fiducial tracker can also provide object location information that is used in the manipulative and configural buffers (described below). Obtaining additional (e.g., semantic) information about an object or person requires declarative retrieval(s).

2.1.2. Manipulative

The Manipulative system parallels Milner and Goodale's (2008) ''perception for action'' dorsal visual pathway. It provides the representations and transformations necessary for the execution of spatially guided action. The system represents objects as metric 3D geons (extracted from a database currently) (Biederman, 1987; Pizlo, 2008) with highly accurate position and orientation, enabling grasping, and other *manipulations*. It has capacity-limited support for the manipulation of those representations through mental rotations and translations (Shepard & Metzler, 1971). This enables the model to engage in complex spatial manipulations and action planning (Milner & Goodale, 2008). The manipulative system

ignores issues of object identity, leaving that to the visual system, tying the two representations through co-occurrence.

In our models below, head orientation was represented using manipulative information. From the functional side, manipulative information is extracted from the environment by a 3D optical flow model to capture a person's 3D head pose in space (Fransen et al., 2009) and a fiducial tracker for determining an object's orientation (Kato et al., 2000). Both the head and object trackers are functional systems that provide accurate metric orientation of different classes of objects (e.g., faces) at a granularity that is consistent with SECS. This information becomes part of the manipulative representations. While the method we are using to get the manipulative information is not cognitively plausible (i.e., we are using current AI computer vision methods), after the information becomes part of our manipulative representation, the rest of the process is cognitively plausible.

2.1.3. Configural

The configural system represents spatial information that is individual to an object and egocentric (as opposed to holistic cognitive maps, O'Keefe & Nadel, 1978). These representations arise from the attending of visual or auditory information in the environment, typically in support of navigation and scene recognition. Configural representations of individual objects are merely egocentric range vectors to the object's defining edges. As with the manipulative system, object identity is left up to the visual system. Specific locations in space are represented as configurations of these individual representations (i.e., two configural representations and the angle between them). In order to maintain spatial consistency across time and space, currently attended representations within working memory are updated automatically by path integration (Presson & Montello, 1994; Rieser, 1989) or consciously by serial mental transformations (Wang, 1999; Wraga, Creem, & Proffitt, 2000).

In the models below, configural information is used to represent whether a specific object is closer or further away from self than another object. Configural information is extracted from the spatial features returned by two tracker systems: the two trackers provide accurate distance and angular information from self at an accuracy and granularity that is consistent with SECS.

2.2. Motor

Traditional ACT-R has a virtual motor system that allows simple, serial, virtual hand movements (e.g., typing, mouse movements). ACT-R/E generalizes this motor system to be applicable to all possible muscle groups (e.g., eyes, head, torso) in parallel. Models are then able to issue motor commands allowing it to move and navigate in virtual and real environments (i.e., when embedded on a robot). ACT-R/E's motor system also permits other modules to monitor efferent motor commands. This allows the configural system to track movements that would be relevant for the updating of configural representations during path integration.

3. Simulator and robot description

Currently, the open-source stage robot simulator (Collett, MacDonald, & Gerkey, 2005) is used to enable data collection and to speed up the model development cycle.

Our current robot platform is the Mobile-Dexterous-Social (MDS) Robot (Breazeal, 2009). The MDS robot neck has 18 DoF for the head, neck, and eyes, allowing the robot to look at various locations in 3D space. Perceptual inputs include two color video cameras and a SR3000 camera to provide depth information. Since people do not have a depth sensor, we do not use that information in the current project.

4. Models of spatial cognition

The remainder of this paper will present two developmental process models that show how children learn spatial competencies along with a match to data. While a match to data is not a perfect measure of cognitive plausibility (Cassimatis, Bello, & Langley, 2008), it can be used to differentiate models. At the least, if a model can show performance and competence as well as a reasonable data fit, it is more plausible (and, to us, preferred), than a model that does not.

The first set of data we model is an experiment by Corkum and Moore (1998) who were interested in what age gaze-following naturally occurs and whether or not children can learn to follow gaze.

4.1. Brief description of Corkum and Moore (1998)

Three age groups (6- to 7-, 8- to 9-, and 10- to 11-month-olds) completed the experiment. The trial started when the infant looked at the experimenter. Each trial consisted of the experimenter looking 90° left or right at one of two toys (see Fig. 1a.)

The experiment consisted of three consecutive phases. In the baseline phase, the experimenter simply looked at a toy. During the baseline phase the toy remained inactive (i.e., did not light up or turn) in order to assess spontaneous gaze-following.

During the shaping phase, regardless of the infant's gaze, the toy that was gazed at by the experimenter lit up and rotated.

During the final testing phase, the toy was activated only if the infant and the experimenter looked at the same toy.

Each head turn was coded as either a target (joint-gaze with the experimenter) or a non-target (the wrong toy was gazed at) response. Random gaze-following would correspond to approximately 50% accuracy, while accurate gaze-following would be greater than 50%.

As Fig. 3 suggests, only 10–11 month infants could reliably follow gaze at baseline. After training, however, both 8–9 month and 10–11 month infants could reliably follow gaze (there was a slight, non-significant increase in gaze-following for the 6–7 month infants).

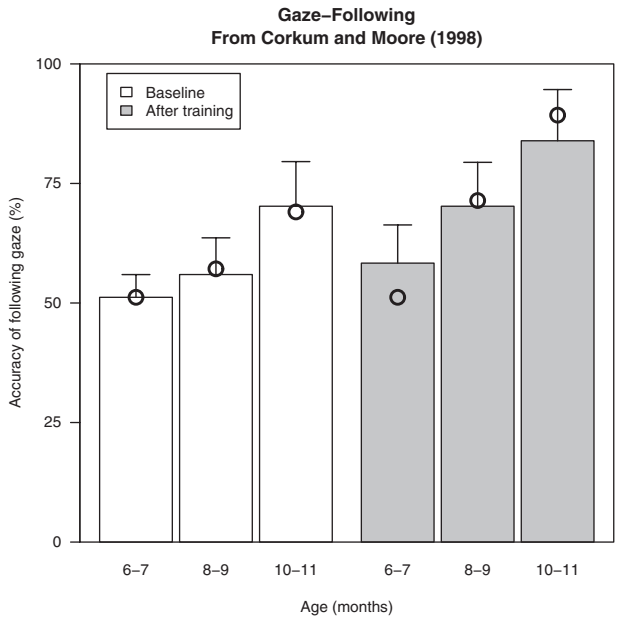

Fig. 3. Experimental data from Corkum and Moore (1998). Bars are experimental data and circles are model data. Error bars are 95% confidence intervals.

These results are consistent with previous data (Corkum & Moore, 1995) showing that gaze-following reliably occurs during the end of the first year: only 10–11 month infants could reliably follow gaze at baseline. Interestingly, however, 8–9 month infants learned to follow gaze in the experimental setting with a modest amount of training.

Corkum and Moore (1998) interpret these data as showing that there are several precursors to gaze-following. First, infants must be mature enough to respond to different spatial locations; they must have some rudimentary spatial ability. Second, infants must be able to learn that an interesting event will occur where the person looks. They further suggest that the adult's head turn cues the infant's attention in the direction of the turn.

5. Model description

An ACT-R/E model was developed that simulates the development of gaze-following.

5.1. High-level description of the gaze-following model

There are five model components that enable gaze-following: the reactive nature of the model; using ACT-R's memory system as a model of habituation; the spatial components; the gaze-following itself; and the utility learning mechanism.

5.1.1. The reactive nature of the model

The model itself is completely bottom-up; there is no goal-directed or top-down action in this model. The model was written in this manner because early gaze-following seems to be emergent rather than goal-directed (Triesch et al., 2006).

5.1.2. Habituation in ACT-R

When the model gazes at any object (person, toy, etc.), it looks at that object until it can recall the object before it attempts to look at a different object. This is an approximation of habituation (Sirois & Mareschal, 2002); several other researchers (Triesch et al., 2006) use an exponential function that is remarkably similar and formally equivalent to ACT-R's model of memory retrieval (Anderson et al., 1998). After the model gazes at and habituates to an object, it starts to look for a new object.

5.1.3. Spatial module

For pure gaze-following, both the visual module and the manipulative modules are needed; configural is not necessary. The visual buffer provides object identification while the manipulative buffer provides the orientation of a particular object. Specifically, the manipulative buffer provides information about what direction a person is facing and presumably gazing.

5.1.4. Gaze-following

Gaze-following was implemented by adding constraints to the visual search mechanism. Gaze-following is a directed visual search along a retinotopic vector. Given a starting point and either an angle or an end point, the visual search will return the location or an object somewhere along that line within some tolerance. This simple mechanism allows the visual system to find candidate objects or obstructions along a gaze. These skills align nicely with Butterworth's earlier developmental stages of gaze-following (Butterworth & Jarrett, 1991).

The current model provides this knowledge as is, though we assume that this rule is created through ACT-R's production compilation process (Taatgen, 2003). Our focus here is on how the different spatial representations can be combined and used, not how the specific rule gets created (e.g., Taatgen & Anderson, 2002; Van Rijn, Van Someren, & Van der Maas, 2003).

5.1.5. Utility learning

ACT-R is able to not only learn new facts and rules but also to learn which rule should fire (called utility learning in ACT-R). ACT-R uses an elaboration of the Rescorla-Wagner learning rule and the temporal-difference (TD) algorithm (Fu & Anderson, 2006). The TD algorithm has been shown to be related to animal and human learning theory (Sutton & Barto, 1981).

Briefly, any time a reward is given (e.g., for infants, a smile from a caregiver), a reward is propagated back in time through the rules that had an impact on the model getting that reward. Punishments may also be given with a similar time course, but no punishments were given in this model.

For all models, we kept most of the ACT-R parameter defaults. The parameters that were changed include the base-level learning (a decay value of .2 instead of the typical default of .5), which allowed for a reasonable habituation timecourse; utility noise (set at a reasonable .5) to allow low-utility productions to occasionally fire; and the utility learning rate (set at .2), which allowed the productions to converge to a stable expected utility within a reasonable period of time (minutes instead of months). Note that the same qualitative behavior of the model emerges across a huge range of parameter values. Utility noise impacts the variability of different production firings (the higher the number, the more variability there is in the final model). Utility learning rate impacts the speed at which the rules converge.

5.2. A sample experimental model run

The first thing that the model does in an experimental trial is to find the caregiver, corresponding to the experimental procedure where the experimenter gets the infant's attention (Corkum & Moore, 1998). The model looks at the caregiver until it has habituated to that person. The caregiver looks at an object in the environment for 7 s or until the model makes a decision about where to look.

When the model is ''young'' it has a favored rule set, which is to locate, attend to, and gaze at an object. The object can be anything in the model's field of view and it is chosen randomly.

If the caregiver is looking at the same object that the model decides to look at, the model is given a small reward. If the caregiver is looking at a different object than the model, no reward is given but the trial is completed and the reward process begins anew.

Even though there is a favored rule to find an object and gaze at it, the gaze-following rule competes with it. The gaze-following rule has a much lower utility when the model is young so it does not get an opportunity to fire very often. However, because there is noise associated with production firings, the gaze-following rule does occasionally get a chance to fire. When the gaze-following rule has a high enough utility to fire, the model attempts to follow the gaze of the caregiver to an object.

The gaze-following production uses manipulative knowledge of the head of the caregiver to determine what direction the caregiver's head is facing. Note also that the model assumes that the eyes are facing the same direction as the head. For the experimental procedure discussed here, this assumption is appropriate, but as children develop (by 1 year) they do differentiate between head pose and where the eyes themselves are gazing (Brooks & Meltzoff, 2002).

With this information, the infant model looks from the caregiver in the direction the head is facing. The model then finds the first available object in that direction, which is consistent with previous research (Butterworth & Jarrett, 1991). The model is again given a small reward. After habituation to that object, the trial ends and the model looks for another object to attend to.

Because the gaze-following production is correct more often than the random production, the gaze-following production slowly gains utility. However, it takes a period of time before

the combination of noise and utility allow the gaze-following production to overtake and eventually become dominant over the random-object production.

5.3. Modeling developmental progress

When the model is young, it has a handful of productions that look around the world. Experience is simulated by concentrating gaze-following learning such that a few minutes is equal to 2 months. We assume, based on previous data (Deák, Wakabayashi, Sepeta, & Triesch, 2004), that the caregivers of very young children (under 6 months) use pointing and object motion to get the infant's attention and that this pointing and object motion may serve as a bootstrapping process for gaze-following (Flom, Deák, Phill, & Pick, 2004). Our model, thus, assumes that gaze-following becomes a viable cue to attention just before 6 months, and that the child receives experience with gaze-following at a constant rate thereafter. Thus, the 6–7 month model was given 80 s of concentrated experience with looking around a simple world at objects and receiving feedback as described in the experimental run. For the 8–9 month model, 3 min of experience were given, and for the 10–11 month model, 6 min of experience were given. Because the rate of learning is dependent entirely on the utility learning rate parameter, learning occurred quite quickly in this model. The utility learning rate could be scaled down substantially to match actual infant learning time. Other researchers have used a very similar approach (Schapiro & McClelland, 2009; Van Rij, Van Rijn, & Hendriks, 2010). In order to do this correctly, however, it would be important to know approximately how many times an infant attempts to follow a gaze or how often an infant receives feedback or the infant found something especially interesting to look at as well as knowledge about the environment (e.g., the number of objects). Other researchers have come to a similar conclusion concerning the importance of learning in gaze-following (Corkum & Moore, 1998; Triesch et al., 2006).

At each age (6–7, 8–9, and 10–11 months), the model was put through the same experimental procedure as Corkum and Moore (1998). Note that the lighting up and rotating of the toy provided a strong reward to the child, which is modeled by joint attention during the training phase of the procedure; no reward was given during the baseline phase, so this was a pure measure of age-related ability.

To provide some match to the experimental procedure, 21 models (corresponding to the 21 participants) were run at each age group. However, to achieve stable results, the model was run 10 times with no utility learning for the baseline and after training conditions. This allowed the model to be tested after different age or experimental related amounts of practice yet maintain stable results.

5.4. Model fit

As is evident in Fig. 3, the model matches the data quite well; $R^2 = .95$ and RMSD = .3. Critically, all model points are within 95% confidence intervals of the data. The model suggests that there is both a qualitative and a quantitative change in the child. Qualitatively, the child reduces the amount of time he or she looks around randomly and

starts to pay attention to other cues in the environment (e.g., the gaze of the caregiver); this non-random behavior provides the child with more rewards. With experience, gaze-following becomes more and more preferred over the random behavior that the model initially preferred. The 6–7 month children did not statistically improve; the model explanation for this is that they simply had not had enough experience yet. Note that with more experimental training, the 6–7 month model would eventually learn to follow gaze. This is a strong prediction of the model: with enough practice, even 6-month-olds should be able to learn to follow gaze. Interestingly, with a modest amount of experimental training, the 8–9 month model also showed improvement. Again the model suggests that the reason for this is that 8–9 month children were at the ''right'' developmental age to take advantage of the concentrated training. This ''right'' age to take advantage of the training is simply enough exposure in the child's (or model's) lifetime. This training allowed productions that occasionally fired during the model's previous experience to be focused and rewarded, which brought their utility to surpass the random behavior they had before the experiment started.

5.5. Discussion of model of Corkum and Moore (1998)

The primary advantage of this model over previous models (e.g., Doniec, Sun, & Scassellati, 2006; Nagai, Hosoda, Morita, & Asada, 2003; Triesch et al., 2006) is that it adds a spatial component integrated into an embodied cognitive architecture, one of the major concerns with previous models (Moore, 2006).

Of the model's five components (reactivity, habituation, the spatial module, gaze-following, and utility learning), three of them are absolutely critical to the success of the model. The reactive nature of the module is a theoretical commitment to modeling young children, though the model could be written using a top-down model. Likewise, habituation is something that has been theoretically proposed and empirically observed, though it is not a critical component to the success of the model. The spatial module, gaze-following, and utility learning, however, are crucial components of the model. The spatial component integrates the spatial aspects of the task while the entire system could not function without the ability to perceive which direction a person is gazing. If the model did not have the spatial representations, it would not be able to use and integrate the pertinent information from the environment. Because the developmental progress is accounted for by utility learning, it also is a necessary part of the model.

The gaze-following model only used visual and manipulative spatial representations. Our theory of spatial cognition should be able to use additional spatial representations when they are needed. To explore this, we focus next on Level 1 visual perspective taking and model an experiment by Moll and Tomasello (2006).

5.6. Brief description of Moll and Tomasello (2006)

Two age groups (18- and 24-month-olds) completed the experiment in one of two conditions. The physical setup was very similar to Corkum and Moore (1998) with the

addition of an occluder in between the experimenter and one of the toys (see Fig. 1b). In the control condition, the experimenter looked back and forth between the occluder and the visible toy, saying, ''Can you give the toy to me?'' In the experimental condition, the experimenter said, ''Where is the other toy? Where is it? I can not find it! Can you give the toy to me?'' A successful trial was coded if the child gave the hidden toy to the experimenter.

As Fig. 4 suggests, only 24-month-olds in the experimental condition could reliably give the experimenter the hidden toy and successfully showed Level 1 visual perspective-taking abilities. In all other cases, a random toy was chosen. Note that there was an oddity in the data: children in the other conditions seemed to have a non-significant preference for giving the unobstructed toy to the experimenter. Moll and Tomasello suggest that the reason for this non-significant preference is that the unobstructed toy provided a clear line of sight across the room.

Moll and Tomasello (2006) interpret these data as showing that 24 month children can perform simple Level 1 visual perspective-taking tasks. They succeeded at this task earlier than most other studies show Level 1 perspective taking because this task used a routine search for objects and did not rely on the child's verbal abilities.

6. Model description

An ACT-R/E model was developed that simulates the development of Level 1 visual perspective taking.

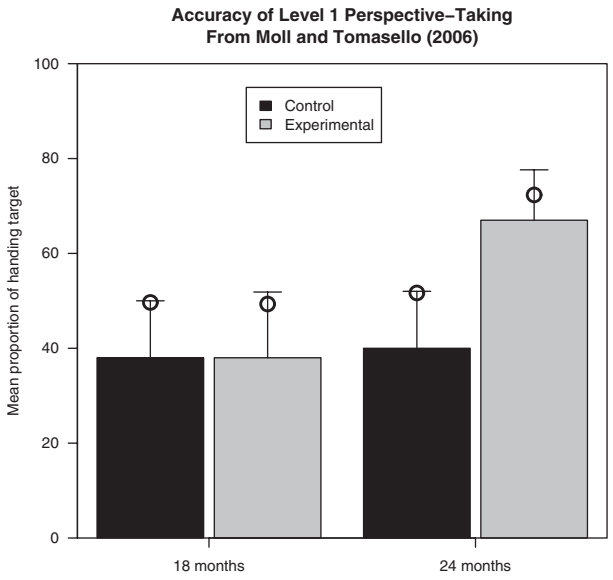

Fig. 4. Experimental data from Moll and Tomasello (2006). Bars are experimental data and circles are model data. Error bars are 95% confidence intervals.

6.1. High-level description of Level 1 visual perspective-taking model

The model for Level 1 visual perspective taking uses most of the same components as the gaze-following model with two major differences. First, the gaze-following model was reactive/bottom-up, while the Level 1 model uses goal-directed/top-down processing. Second, the gaze-following model did not use configural representations, but Level 1 needs configural knowledge to be able to determine whether the occlusion is between the experimenter and one of the toys.

This was a zero-free-parameter fit; the model kept the same parameter values as the gaze-following model.

6.2. A sample experimental model run

The first thing that the model does in an experimental trial is to find a caregiver. In the experimental condition, the model has the goal to find a toy that is not visible from the experimenter's location. This goal comes directly from the experimenter not being able to find one of the toys. We assume that this goal is given through natural language, though we do not perform natural language processing in the current model.

When the model is ''young'' it has a favored rule set, which is to locate, attend to, gaze at, and get an object. The object can be anything in the model's field of view and it is chosen randomly.

If the model selects the object that is, in fact, hidden from the caregiver, the model gets a small reward. If the model selects the object that is visible to both caregiver and the model, the model receives a smaller reward. Note that the model is not told directly whether the toy is hidden or not from the caregiver; it must derive this information from the spatial positions of the toy and the caregiver.

As in the gaze-following model, we assume that the Level 1 visual perspective-taking rule is created by production compilation as in previous ACT-R developmental models (e.g., Taatgen & Anderson, 2002; Van Rijn et al., 2003). The Level 1 visual perspective-taking rule competes with the default rule to simply give the caregiver a toy. Again, the perspective-taking rule starts off with a low initial utility and eventually overtakes the default set of rules. Like the gaze-following model, a wide range of parameter values provides the same qualitative behavior.

In this model, visual perspective taking requires a series of steps to succeed. After the model has found a toy, a series of spatial operations need to be made to determine whether that toy can be seen by the caregiver. The first check is to perform a gaze-following operation: are there any possible objects that could be occluding the view of the caregiver to the toy? A possible occlusion is simply an object between the caregiver and the toy: this check requires no distance information and is retinotopically between the caregiver and the toy.

If there are no possible occlusions, the model looks for a different toy and starts the process anew. If there is a possible occlusion, the object's distance is checked to see if it is an actual occlusion rather than a possible occlusion. Two checks using configural knowledge need to

be made at this stage. First, the object must be closer to the model than the caregiver is; if the object is further away than the adult, it is not occluding the caregiver's view of the toy. Second, the object must be further away from the model than the toy is; if the object is closer than the toy, it is not occluding the caregiver's view of the toy. The model uses its configural representations (egocentric range vectors) to make these comparisons. Note that in order to be a true obstruction, the leftmost part of the object must be further to the left than the leftmost part of the caregiver/toy or the rightmost part of the object must be further to the right than the rightmost part of the caregiver/toy. This left/right check is also made using configural knowledge. Finally, the object must be at least as big as the toy in order to occlude it. If, after these spatial checks and comparisons, the object is still considered an occlusion, the toy that was chosen is given to the caregiver. If the appropriate toy is given to the caregiver, the model receives a small reward. Note that perspective taking is accomplished through an entire series of productions, not a unitary ''perspective-taking'' rule. Also note that the model is able to give the caregiver the jointly visible toy if asked (e.g., ''Give me the toy.'').

This process by necessity requires some top-down reasoning. First, the model must keep track of which toy(s) it has checked. Second, the model must be able to store the toy it is working on while verifying that the caregiver can (or cannot) see it, perform additional configural operations on different objects in the environment, and make a decision. These storage operations as well as the series of steps that must occur mandate using a top-down approach.

As with gaze-following, Level 1 perspective taking is more successful than the default productions and over time, the Level 1 production becomes dominant over the default productions. Eighteen model runs (corresponding to eighteen participants) were executed at each age group/condition.

6.3. Model fit

As is evident in Fig. 4, the model matches the data quite well at a qualitative level; $R^2 = .99$ but less well at a quantitative level; RMSD = 10.2. The reason for the poorer quantitative fit is that we chose not to model the preference for choosing a toy that had a clear line of sight across the room because there is no architectural reason this preference would have manifested itself in this situation and not others (e.g., the Corkum and Moore dataset modeled earlier) and the fact that the difference was not theoretically relevant to Level 1 perspective taking. If this result is replicable and becomes important to the development of spatial competence or Level 1 visual perspective taking, it will need to be accounted for in future models. Even with the data anomaly, all model points are within 95% confidence intervals of the data. We interpret this fit as overall positive. First, we are able to capture the qualitative aspects of this task quite well. Given that there are no other existing quantitative models of this task, the model seems quite reasonable as a process description.

6.4. Discussion of Level 1 visual perspective-taking model

We described an embodied model of Level 1 visual perspective taking that built on our gaze-following model, using many of the same capabilities that gaze-following requires.

Our model matches data from an experiment that showed that even 24 month children could demonstrate Level 1 perspective taking. While the gaze-following model used primarily manipulative spatial representations, Level 1 visual perspective taking required configural spatial representations, primarily for determining whether or not an object is actually occluding a person's view.

Similar to the gaze-following model, the Level 1 model requires spatial knowledge and utility learning. This model used gaze-following to find an initial toy, but that was not a fundamental aspect of the model. Unlike the gaze-following model, Level 1 visual perspective taking requires some amount of goal-directed reasoning.

7. Embodied modeling

Both the gaze-following model and the Level 1 perspective-taking model were run on an embodied platform (our robot). The testing environment was set up in the same manner as Corkum and Moore (1998) and Moll and Tomasello (2006). Note, however, that the models' behaviors are not dependent upon the specific spatial configurations, objects, or individuals involved.

One interesting aspect of the occlusion robot demo was that if the occluder was too small and did not prevent the caregiver from seeing the ''hidden'' toy, the model would assume that the caregiver could see both objects and choose a toy at random. These conditions were not manipulated in the experiments for experimental control reasons, but the robot model is able to robustly deal with a wide range of differences in scene and objects and maintain the same behavioral performance.

Movies are available at http://www.nrl.navy.mil/aic/iss/aas/CognitiveRobotsVideos.php.

8. General discussion

8.1. SECS

We have presented SECS, a neuro-computational framework for spatial cognition with three spatial representations (visual, manipulative, and configural). SECS supports the coordination of multiple representations and frames of reference that can be created by different modalities (Xing & Andersen, 2000). SECS uses amodal, egocentric, and updatable representations (Avraamides, Loomis, Klatzky, & Golledge, 2004). Even though it is fundamentally egocentric, we suggest that the representations used by SECS can be used to account for a wide variety of spatial cognition tasks.

8.2. The development of spatial competencies

Our work provides support for the hypothesis put forth by Newcombe and Huttenlocher (2000) that development consists of changes in the importance attached to different types of

spatial information. In our models of both gaze-following and Level 1 visual perspective taking, the model starts off with access to spatial information about objects in the world, but it does not start off using it appropriately. With practice and feedback, the model learns to use some spatial information, showing developmental competence. Both these models start off with a great deal of spatial competency, but they are not put together the right way. For example, Level 1 visual perspective taking can follow gaze, but gaze-following is not the critical spatial competency that needs to be used; obstruction and perspective taking are much more important. Experience and feedback lead to the correct selection of spatial competencies in service of different tasks. Also note that the models presented here work with a very wide range of parameters; the specific parameter values fit the data nicely, but the qualitative behavior of the models works across most parameter values.

This work takes the proposal put forth by Spelke that infants have core spatial knowledge and combined that with Newcombe and Huttenlocher's suggestion that spatial competence develops by reweighting the different spatial information as it is needed and comes into conflict. Finally, we have developed our models on an embodied platform.

8.3. Embodied spatial cognition

Embodied representations provide a strong approach to studying spatial cognition. By embodying cognitive models, the importance of spatial cognition becomes obvious: different aspects of spatial cognition are manifested in even seemingly simple tasks like gaze-following or perspective taking. Our models of spatial cognition are a strong example of this embodied view.

Note

1. Additional technical detail about SECS is available at http://anthonymharrison.com/blog/secs/.

Acknowledgments

This work was supported by the Office of Naval Research under funding documents N0001408WX30007, N000141WX20474, and N0001411WX20407 to JGT.

References

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York: Oxford University Press.

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language*, *38*, 341–380.

Anderson, J. R., Qin, Y., Jung, K. J., & Carter, C. S. (2007). Information-processing modules and their relative modality specificity. *Cognitive Psychology*, *54*(3), 185–217.

Avraamides, M. N., Loomis, J. M., Klatzky, R. L., & Golledge, R. G. (2004). Functional equivalence of spatial representations derived from vision and language: Evidence from allocentric judgments. *Journal of Experimental Psychology: Human Learning, Memory, & Cognition*, *30*, 801–814.

Baron-Cohen, S. (1995). The eye direction detector (EDD) and the shared attention mechanism (SAM): Two cases for evolutionary psychology. In C. Moore & P. Dunham (Eds.), *Joint attention: Its origins and role in development* (pp. 41–59). Hillsdale, NJ: Lawrence Erlbaum.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115–147.

Breazeal, C. (2009). MDS robot. [accessed May 1, 2009]. Available at http://robotic.media.mit.edu/projects/robots/mds/overview/overview.html.

Brooks, R. A., & Mataric, M. J. (1993). Real robots, real learning problems. In J. Connell & S. Mahadevan (Eds.), *Robot learning* (pp. 193–213). Dordrecht, Netherlands: Kluwer Academic Press.

Brooks, R., & Meltzoff, A. N. (2002). The Importance of eyes: How infants interpret adult looking behavior. *Developmental Psychology*, *38*, 958–966.

Butterworth, G., & Jarrett, N. (1991). What minds have in common is space: Spatial mechanisms serving joint visual attention in infancy. *British Journal of Developmental Psychology*, *9*, 55–72.

Cassimatis, N. L., Bello, P., & Langley, P. (2008). Ability, breadth and parsmiony in computational models of higher-order cognition. *Cognitive Science*, *32*(8), 1304–1322.

Collett, T. H., MacDonald, B. A., & Gerkey, B. P. (2005). Player 2.0: Toward a practical robot programming framework. In *Proceedings of the Australasian Conference on Robotics and Automation (ACRA 2005)*. Sydney: Australia.

Corkum, V., & Moore, C. (1995). Development of joint visual attention in infants. In V. Corkum & C. Moore (Eds.), *Joint attention: Its origins and role in development* (pp. 61–83). Hillsdale, NJ: Lawrence Erlbaum.

Corkum, V., & Moore, C. (1998). The origins of joint visual attention in infants. *Developmental Psychology*, *34*(1), 28–38.

Deák, G. O., Wakabayashi, Y., Sepeta, L., & Triesch, J. (2004). Development of attention-sharing from 5 to 10 months of age in naturalistic interactions. In *Proceedings from International Conference on Infancy Studies*, Chicago, IL.

Dehaene, S., Izard, V., Pica, P., & Spelke, E. (2006). Core knowledge of geometry in an Amazonian indigene group. *Science*, *311*(5759), 381–384.

Doniec, M. W., Sun, G., & Scassellati, B. (2006). Active learning of joint attention. In *IEEE-RAS International Conference on Humanoid Robotics, Genova, Italy*.

Flavell, J. H. (1999). Cognitive development: Children's knowledge about the mind. *Annual Review of Psychology*, *50*, 21–45.

Flom, R., Deák, G. O., Phill, C. G., & Pick, A. D. (2004). Nine-month-olds' shared visual attention as a function of gesture and object location. *Infant Behavior and Development*, *27*(2), 181–194.

Fransen, B., Hebst, E., Harrison, A., & Trafton, J. G. (2009). 3D position and pose tracking. In *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*. St. Louis, MO.

Fu, W.-T., & Anderson, J. R. (2006). From recurrent choice to skill learning: A model of reinforcement learning. *Journal of Experimental Psychology: General*, *135*(2), 184–206.

Gunzelmann, G., & Lyon, D. R. (2007). Mechanisms for human spatial competence. In T. Barkowsky, M. Knauff, G. Ligozat, & D. Montello (Eds.), Spatial cognition V: Reasoning, action, interaction (pp. 288–307). Berlin: Springer-Verlag.

Harrison, A. M. (2007). *Online or offline? Exploring working memory contraints in spatial updating*. Unpublished doctoral dissertation. University of Pittsburgh.

Harrison, A. M. & Schunn, C. D. (2002). ACT-R/S: A computational and neurologically inspired model of spatial reasoning. In W. D. Gray & C. D. Schunn (Eds.), *Proceedings of the Twenty Fourth Annual Meeting of the Cognitive Science Society* (p. 1008). Fairfax, VA: Lawrence Erlbaum Associates.

Kato, H., Billinghurst, M., Poupyrev, I., Imamoto, K., & Tachibana, K. (2000). Virtual object manipulation on a table-top AR environment. In *IEEE and ACM International Symposium on Augmented Reality* (pp. 111–119). Los Alamitos, CA: IEEE CS Press.

Klatzky, R. L. (1998). Allocentric and egocentric spatial representations: Definitions, distinctions, and interconnections. In C. Freksa, C. Habel & K. F. Wender (Eds.), *Spatial cognition: An interdisciplinary approach to representing and processing spatial knowledge* (pp. 1–17). New York: Springer-Verlag.

McNamara, T. P., & Shelton, A. L. (2003). Cognitive maps and the hippocampus. *Trends in Cognitive Sciences*, *7*(8), 333–335.

Milner, A. D., & Goodale, M. A. (2008). Two visual systems re-viewed. *Neuropsychologia*, *46*(3), 774–785.

Moll, H., & Tomasello, M. (2006). Level 1 perspective-taking at 24 months of age. *British Journal of Developmental Psychology*, *24*, 603–613.

Moore, C. (2006). Modeling the development of gaze following needs attention to space. *Developmental Science*, *9*, 149–150.

Nagai, Y., Hosoda, K., Morita, A., & Asada, M. (2003). A constructive model for the development of joint attention. *Connection Science*, *15*(4), 211–229.

Newcombe, N. S., & Huttenlocher, J. (2000). *Making space*. Cambridge, MA: MIT Press.

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford, England: Clarendon Press.

Pick, H. L., & Rieser, J. J. (1982). Children's cognitive mapping. In M. Potegal (Ed.), *Spatial orientation: Development and physiological foundations* (pp. 107–128). New York: Academic Press.

Pizlo, Z. (2008). *3D shape: Its unique place in visual perception*. Cambridge, MA: The MIT Press.

Presson, C. C., & Montello, D. R. (1994). Updating after rotational and translational body movements: Coordinate structure of perspective space. *Perception*, *23*, 1447–1455.

Previc, F. H. (1998). The neuropsychology of 3-D space. *Psychological Bulletin*, *124*(2), 123–164.

Rieser, J. J. (1989). Access to knowledge of spatial structure at novel points of observation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(6), 1157–1165.

Scaife, M., & Bruner, J. S. (1975). The capacity for joint visual attention in the infant. *Nature*, *253*, 265–266.

Schapiro, A. C., & McClelland, J. L. (2009). A connectionist model of a continuous developmental transition in the balance scale task. *Cognition*, *110*(3), 395–411.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*, 701–703.

Sirois, S., & Mareschal, D. (2002). Models of habituation in infancy. *Trends in Cognitive Sciences*, *6*(7), 293–298.

Spelke, E. S. (2003). What makes us smart? Core knowledge and natural language. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language* (pp. 277–311). Cambridge, MA: MIT Press.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*, 89–96.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, *88*, 135–140.

Taatgen, N. A. (2003). Production compilation: A simple mechanism to model complex skill acquisition. *Human Factors*, *45*(1), 61–76.

Taatgen, N. A., & Anderson, J. R. (2002). Why do children learn to say ''broke''? A model of learning the past tense without feedback. *Cognition*, *86*(2), 123–155.

Triesch, J., Teuscher, C., Deak, G. O., & Carlson, E. (2006). Gaze following: Why (not) learn it? *Developmental Science*, *9*(2), 125–147.

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale & R. J. W. Mansfield (Eds.), *Analysis of visual behavior* (pp. 549–586). Cambridge, MA: MIT Press.

Van Rij, J., Van Rijn, H., & Hendriks, P. (2010). Cognitive architectures and language acquisition: A case study in pronoun comprehension. *Journal of Child Language*, *37*(3), 731–766.

Van Rijn, H., Van Someren, M., & Van der Maas, H. (2003). Modeling developmental transitions on the balance scale task. *Cognitive Science*, *27*(2), 227–257.

Wang, R. F. (1999). Representing a stable environment by egocentric updating and invariant representations. *Spatial Cognition and Computation*, *1*(4), 431–445.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*, 625–636.

Wraga, M., Creem, S. H., & Proffitt, D. R. (2000). Updating displays after imagined object and viewer rotations. *Journal of Experiment Psychology: Learning, Memory, and Cognition*, *26*(1), 151–168.

Xing, J., & Andersen, R. A. (2000). Models of the posterior parietal cortex which perform multimodal integration and represent space in several coorinate frames. *Journal of Cognitive Neuroscience*, *12*, 601–614.