Title
A reinforcement learning model of gaze following

Author
Jasso, Hector

Publication Date
2007

UNIVERSITY OF CALIFORNIA, SAN DIEGO

A Reinforcement Learning Model of Gaze Following

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Computer Science and Cognitive Science

by

Hector Jasso

Committee in charge:

Professor Garrison W. Cottrell, Co-chair
Professor Jochen Triesch, Co-chair
Professor Serge J. Belongie
Professor Charles Elkan
Professor Jeffrey L. Elman

2007

The dissertation of Hector Jasso is approved, and it is acceptable in quality and form for publication on microfilm:

Co-Chair

Co-Chair

University of California, San Diego

2007

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to gratefully acknowledge the enthusiastic supervision of Professors Jochen Triesch and Gary Cottrell during this work.

I thank members of the following teams for stimulating discussions, technical assistance, office space, bagels, music sharing, pool games, and expeditions in search of good food:

- Modeling the Emergence of Shared Attention (MESA) team, especially Professor Gedeon Deák and Dr. Christof Teuscher, who participated closely in the development of the model and were co-authors in the related publications,
- Gary's Unbelievable Research Unit (GURU),
- Artificial Intelligence Laboratory (AI lab),
- Complex Systems and Cognition Laboratory (CSC lab),
- Cyberinfrastructure Laboratory for Environmental Observing Systems (CLEOS), at the San Diego Supercomputer Center (SDSC), where I worked as a research assistant during my studies. In particular, Dr. Tony Fountain, director of CLEOS, for his patience and support. Thanks to Dr. Chaitan Baru, Division Director of Science Research and Development at SDSC, for his support.

To my parents and family, for their continuous support and encouragement.

To Natasha, for lovingly watching over me, and for being a firm believer of the mirror neuron.

The text of chapters 3 and 4, in part, is a reprint of the material in:

- H. Jasso, J. Triesch, C. Teuscher, and G. O. Deák. A reinforcement learning model explains the development of gaze following. In *Proceedings of the Seventh International Conference on Cognitive Modeling (ICCM 2006)*, Trieste, Italy, April 2006.

The dissertation author was the primary investigator and author of this paper.
The text of chapter 5, in part, is a reprint of the material in:

- H. Jasso, J. Triesch, and G. O. Deák. Using eye direction cues for gaze following - a developmental model. In *Proceedings of the Fifth International Conference on Development and Learning (ICDL'06)*, Bloomington, IN, May 2006.

The dissertation author was the primary investigator and author of this paper.
The text of chapter 6, in part, is a reprint of the material in:

- J. Triesch, H. Jasso, and G. O. Deák. Emergence of mirror neurons in a model of gaze following. In *Proceedings of the Fifth International Conference on Development and Learning (ICDL'06)*, Bloomington, IN, May 2006.

- J. Triesch, H. Jasso, and G. O. Deák. Emergence of mirror neurons in a model of gaze following. *Adaptive Behavior*, (in press).

The dissertation author was a co-author of these papers.
The text of chapter 7, in part, is a reprint of the material in:

- H. Jasso and J. Triesch. A virtual reality platform for modeling cognitive development. In *Proceedings of the Third International Conference on Development and Learning (ICDL'04)*, La Jolla, CA, October 2004.

The dissertation author was the primary investigator and author of this paper.

VITA

1988	Bachelor of Science, Instituto Tecnológico y de Estudios Superiores de Monterrey (ITESM)
1991	Master of Science, University of Edinburgh
2007	Doctor of Philosophy, University of California, San Diego

PUBLICATIONS

H. Jasso and J. Triesch. A virtual reality platform for modeling cognitive development. In *Proceedings of the Third International Conference on Development and Learning (ICDL'04)*, La Jolla, CA, October 2004.

H. Jasso, J. Triesch, C. Teuscher, and G. O. Deák. A reinforcement learning model explains the development of gaze following. In *Proceedings of the Seventh International Conference on Cognitive Modeling (ICCM 2006)*, Trieste, Italy, April 2006.

H. Jasso, J. Triesch, and G. O. Deák. Using eye direction cues for gaze following - a developmental model. In *Proceedings of the Fifth International Conference on Development and Learning (ICDL'06)*, Bloomington, IN, May 2006.

J. Triesch, H. Jasso, and G. O. Deák. Emergence of mirror neurons in a model of gaze following. In *Proceedings of the Fifth International Conference on Development and Learning (ICDL'06)*, Bloomington, IN, May 2006.

J. Triesch, H. Jasso, and G. O. Deák. Emergence of mirror neurons in a model of gaze following. *Adaptive Behavior*, (in press).

ABSTRACT OF THE DISSERTATION

A Reinforcement Learning Model of Gaze Following

by

Hector Jasso
Doctor of Philosophy in Computer Science and Cognitive Science
University of California, San Diego, 2007
Professor Garrison W. Cottrell, Co-Chair
Professor Jochen Triesch, Co-Chair

Gaze following is defined as the redirection of one's visual attention to match the object of attention of another person. It is a basic mechanism resulting in *joint attention*, the coordination of attention between persons towards an external object. Joint attention in turn is foundational for skills such as imitation, word learning, the interpretation of novel events, and *theory-of-mind*, the understanding of others' beliefs, desires, and intentions.

Gaze following develops within the first 18 months of age, with a gradual improvement in the discrimination of the referred target, the incorporation of eye direction cues in addition to the earlier use of head direction cues, and the appearance of gaze following to out-of-view targets. This development has been interpreted as the gradual incorporation of qualitatively different mechanisms that improve gaze following. For example, it is believed that behind this development is a transition from a simple following of cues (attentional gaze following) into an understanding of others as agents with their own focus of attention (mentalist gaze following). Others see this development as the product of a gradual introduction of different attentional mechanisms that improve spatial aspects of gaze following. This dissertation presents a new computational model of gaze following based on *reinforcement learning*, a biologically plausible learning algorithm. The model replicates the developmental trajectory of gaze following as measured experimentally in key studies. This unifies attentional and mentalist interpretations of gaze following into a learning account. It also offers a parsimonious, single-mechanism account of the improvement of spatial aspects of gaze following.

The dissertation also explores, in the context of the gaze following model: the developmental origins of mirror neurons and their role in imitative behavior; why gaze following tends to develop less in individuals with autism spectrum disorders; and how top-down influences can be seamlessly incorporated into bottom-up visual search. A realistic virtual reality environment platform, built to test gaze following and other cognitive development phenomena, is described.

I

Introduction

I.A Gaze Following as Social Understanding

I.A.1 Social understanding

There has always been a big interest in *social understanding*, the cognitive ability through which we make sense of, predict, and manipulate the behavior of others (Barresi and Moore, 1996; Premack and Woodruff, 1978; Astington, Harris, and Olson, 1988; Perner, 1991; Dennet, 1987; Baron-Cohen, Tager-Flusberg, and Cohen, 2000; Whiten, 1991; Astington, 1993; Flavell, 2000; Frye and Moore, 1991; Moore, 2006; Wellman, 1991; Whiten and Perner, 1991). Following are some examples of the importance of social understanding:

- Without social understanding we would stare at their finger of someone pointing instead of following their pointing. This understanding of referential communication is important for word learning: we teach others by pointing at objects while naming them, expecting them to form the association between words and objects.

- Without social understanding, we would not see a person fetching a ladder as part of a plan to reach a light bulb. Instead, we would see them as simply heading towards the ladder's location. This understanding of others' actions as goal-directed is a crucial component of imitation, where we replicate their goals (reaching a light bulb) by the same means (fetching a ladder) or similar ones (e.g. reaching a box).

- The adaptive value of social understanding can manifest itself in complex sequences, involving and understanding of others' social understanding: We can point to an object (e.g. a rock) in order to drive another's attention away from another object we don't want them to see (e.g. food, if we are competing for it). It is believed that this "outsmarting" of members of the social group, by either anticipating their actions or manipulating their expectations through actions, escalated in primates and resulted in present human intelligence (Byrne and Whiten, 1988; Humphrey, 1976; Jolly, 1966).

I.A.2 Joint visual attention and gaze following

A cornerstone of social understanding is *joint attention*, the ability to coordinate one's attention with the attention of others so that both attend to the same (physical or mental) object (Moore and Dunham, 1995). With the acquisition of joint attention abilities, the human infant transforms its earlier *dyadic* interactions (interactions between it and an adult) into *triadic* interactions (now including objects around the two) (Werner and Kaplan, 1963; Trevarthen and Hubley, 1978; Brazelton, Koslowski, and Main, 1974; Adamson, 1995).

Joint *visual* attention (also called *deictic gaze*) is a type of joint attention where the *visual* attention of two individuals is directed to the same object. This thesis focuses on a basic mechanism that results in joint visual attention: *gaze following*, defined as redirecting one's visual attention based on the other person's head or eye direction (see Fig. I.1). Gaze following precedes other mechanisms that also result in joint visual attention such as pointing and verbal requests (Carpenter, Nagell, Tomasello, Butterworth, and Moore, 1998).

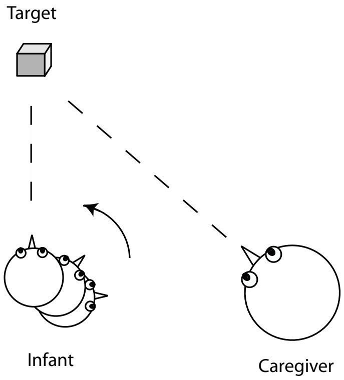


Figure I.1: An illustration of gaze following.

I.B Experimental Measurements of Gaze Following

Michael Scaife and Jerome Bruner (1975) were the first to measure the emergence of gaze following under strict experimental conditions. In their setup, infant and experimenter sit facing each other (similar to Fig. I.1, but without an object). After the experimenter grabs the infant's attention, he/she looks either to the right or to the left, as if looking at an object, for about six seconds. It is then measured whether the infant follows the experimenter's gaze. Results showed that infants as young as three months have a basic capacity for gaze following, a finding that went against prevalent notions of egocentric infants.

This setup has been used extensively to study further aspects of gaze following, leading to the following findings, among others:

- Younger infants do not follow gaze to locations outside their field of view, but older infants do (Butterworth and Cochran, 1980; Butterworth and Jarrett, 1991; Deák, Flom, and Pick, 2000).

- Younger infants will mainly use the adult's head direction as a cue to the location of interesting objects, but older infants will use both head and eye direction cues (Butterworth and Jarrett, 1991; Corkum and Moore, 1995; Brooks and Meltzoff, 2002, 2005).

- Younger infants have trouble disregarding distracter objects which are not on the caregiver's line of sight (Butterworth and Cochran, 1980; Butterworth and Jarrett, 1991; Morissette, Ricard, and Dcarie, 1995).
- Autistic individuals exhibit less gaze following behavior (Leekam, Hunnisett, and Moore, 1998; Mundy, Sigman, and Kasari, 1990).

I.C Nativistic vs. Learning Accounts

Accounts of the developmental origins of gaze following go from the nativistic to learning-based. While in nativistic accounts the skill is said to be genetically specified, developing as the underlying neurological substrate matures, in learning accounts it is seen as acquired through a general learning mechanism that develops through interactions with the world.

Nativistic accounts are easier to propose: one only has to posit the skill as genetically encoded and describe how the different stages in the development of the skill are related to maturational stages in the neurological substrate said to support the skill. Learning accounts, on the other hand, must show how experiences with the environment can drive the model from its original state to one where the skill is mastered. Learning accounts are preferable to nativistic ones because the final state of the mechanism does not have to be genetically encoded and can instead be learned. This is important given the limited genetic information that can be encoded in DNA (Elman et al., 1996).

I.D Outline

Chapter II of this dissertation reviews existing nativistic and learning-based accounts of gaze following, both theoretical as well as computational and robotic. Chapter III proposes a novel account of gaze following based on reinforcement learning (Dayan and Abbott, 2001). In chapter IV the model is used to replicate the experimental observations done by George Butterworth and colleagues

(Butterworth and Cochran, 1980; Butterworth and Jarrett, 1991; Butterworth and Grover, 1988, 1990; Butterworth, 1991, 1995) that led them to believe that behind gaze following are three nativistic perceptual mechanisms that come online in a timely fashion. The reinforcement learning model proposed presents a more parsimonious account, through a single learning mechanism that progresses through the same stages associated to Butterworth's three mechanisms. And since the model does not explicitly represent the caregiver's focus of attention, it helps support Butterworth's intuition that a sophisticated understanding of others' attentional states is not necessary for gaze following to happen. Another set of experiments that are accounted for by the model are described in chapter V. These have to do with the gradual incorporation of eye direction cues to an earlier use of head direction cues (Corkum and Moore, 1995; Brooks and Meltzoff, 2002, 2005; Caron, Butler, and Brooks, 2002), a developmental change that has been interpreted as a shift from a rudimentary form of gaze following to "true" gaze following. By capturing this transition, the model unifies these two forms of gaze following under a single learning-based account. It also shows that the incorporation of eye cues does not require an explicit understanding of other's attention, as is usually assumed. Chapter VI shows how an aversion to social stimuli such as faces and eyes can account for a delay in the development of gaze following, as observed in many autistic individuals. The chapter also shows how, as the model learns to follow gaze, some of its elements develop the same characteristics as mirror neurons (Rizzolatti and Craighero, 2004), which have been considered as a possible neurological underpinning of imitation, another social understanding skill. Finally, chapter VII shows how the model is being extended using realistic virtual reality simulations to explore other aspects of gaze following, thus adding validity to the premise that gaze following can be learned in the real world.

II

Previous Models of Gaze Following

This chapter reviews the different accounts of gaze following, with an emphasis on their categorization as either nativistic or learning-based. The first section reviews theoretical accounts of gaze following, while the second section reviews computational and robotic models of gaze following.

II.A Theoretical Accounts of Gaze Following

II.A.1 Butterworth's three mechanisms

George Butterworth and colleagues (Butterworth and Cochran, 1980; Butterworth and Grover, 1988, 1990; Butterworth and Jarrett, 1991; Butterworth, 1995) presented one of the first nativistic accounts of gaze following. Their motivation was to show that gaze following behavior in infants could be explained without needing to attribute them a complex understanding of the attentional state of others. Instead, they proposed that a set of perceptual mechanisms were sufficient to explain gaze following: In its most basic form of gaze following, the infant simply turns to the side of the room indicated by the mother's gaze, scanning until it finds an interesting object. This would be triggered by seeing the caregiver in a particular pose, along with the presence of an object within its field of view. This stage was named *ecological* because it is the environment (the object's saliency) that completes gaze following.

In this stage, however, the infant does not necessarily stop at the object that the caregiver looks at. Instead, there is a chance that as it scans to the left or the right, other objects not along the caregiver's line of sight might capture its attention instead. Butterworth believed that this was later solved as the infant incorporated into its ecological mechanism a new "geometric compensation mechanism", allowing it to distinguish which objects are located along the caregiver's line of sight, thus avoiding distracters. With this mechanism, the infant is said to have advanced to the *geometric* stage.

Another deficiency in gaze following gets resolved with age: At first, infants will not turn when the object being looked at by the adult is located behind the infant and thus outside its field of view. Butterworth suggests that this gets resolved as the infant gains the ability to represent objects that are out of its sight, that is, once it knows about object permanence (the fact that objects that are not visible still exist). With this mechanism, the infant is said to have advanced to the *representational* stage.

These three mechanisms (ecological, geometric, and representational) do not necessarily replace each other, but instead operate at the same time, newer ones superimposed on older ones. The geometric mechanism supplements the ecological mechanism by giving information about what objects are to be excluded from visual search, and the representational mechanism acts in cases where the adult turns but there are no objects within the infant's field of view to trigger either the ecological or the geometric mechanisms. The geometric mechanism does not completely override the ecological mechanism, though: the infant might still prefer a very salient distracter object over the one looked at by the caregiver.

Butterworth's explanation is nativistic: he described his mechanisms in terms of Piagetian sensorimotor functions, which involve some learning but rely heavily on maturational processes for their development. For example, in Butterworth and Grover (1988), he described the geometric mechanism as arising from "a cognitive developmental process of the Piagetian type", linking it with Piaget's (Piaget and Inhelder, 1948/1956) description of 'invisible displacements' where infants can infer unseen trajectories between different positions in space. He also compared the representational mechanism with Piaget's (Piaget and Inhelder, 1948/1956) description of infants representing space as containers. Similarly, he described the origin of protoimperative pointing (pointing in order to have the caregiver reach an object for him/her), another joint attention skill, as probably related to frontal lobe maturation, implicated in other types of reasoning which require a solution by indirect means (Diamond, 1988).

II.A.2 Leslie's ToBy, ToM-1, and ToM-2

Alan Leslie (1994) presented a nativistic account of gaze following within a framework that included theory-of-mind. He proposed three hierarchically arranged processing subsystems that deal with three main classes of world properties in the agency domain (as opposed to, say, the language, vision, or number domains). These are: a *Theory of Body* mechanism (*ToBy*), and two *Theory of Mind* mechanisms (*system-1* and *system-2 ToMM*). Through ToBy, which develops at around 3 to 4 months, the infant understands mechanical agency: physical causality in a mechanical sense. The two ToMM mechanisms, in turn, deal with the "intentional" properties of agents: system-1 ToMM is used to detect primitive actions such as approach, avoidance, and escape; system-2 ToMM represents attitudes to the truth of propositions (beliefs, wants, pretense) and goal-directedness. Gaze following is attributed by Leslie to the workings of the system-1 ToMM. He believes that system-1 ToMM appears at around 6 months of age based on Butterworth's experiments of the emergence of gaze following (Butterworth and Cochran, 1980; Butterworth and Jarrett, 1991). According to Leslie, these three subsystems become operational once the maturational status of the appropriate neural circuits is achieved. They can develop in sequence or in parallel, or in a combination of the two, where they begin their development in sequence, each one continuing their development in parallel. Although their basic workings are product of innate endowments, these subsystems develop further through learning and based on the availability and quality of its inputs.

II.A.3 Baron-Cohen's ID, EDD, SAM, and ToMM

Simon Baron-Cohen's nativistic theory-of-mind model (1995a; 1995b) is based on four modules: the Intentionality Detector (ID) module, the Eye Direction Detector (EDD) module, the Shared Attention Mechanism (SAM) module, and the Theory-of-Mind Mechanism (ToMM) module. These modules are hierarchical: SAM requires ID and EDD, and ToMM requires SAM.

According to Baron-Cohen, gaze following happens in two stages. First, the EDD module detects eyes and builds representations of eye behavior, called *dyadic representations*. These representations specify the presence of two entities standing in relation to each other. Then the SAM module identifies when the individual and another person are attending to the same thing, building triadic representations based on the EDD's dyadic representations.

Baron-Cohen's idea of the EDD module is based on evolutionary psychology (Cosmides et al., 1992), according to which the architecture of the human mind is inherited and shaped by evolutionary processes. In particular, he cites evidence that specific cells in the Superior Temporal Sulcus (STS) of the monkey brain respond to different perspective view of the head and to direction of gaze (Perret et al., 1985, 1990), with lesions in the STS causing an impairment in this ability (Campbell et al., 1990). Baron-Cohen gives few details about how the complex dyadic representations processed by the EDD module could be implemented. And although the adaptive value of the SAM module (its communicative function and the coordination of individuals) and the ToM module (implementation of social intelligence) makes them also valuable candidates for innate specification through inheritance, no particular brain structures are cited for them.

II.A.4 Meltzoff and Brooke's "Like Me" hypothesis

Andrew Meltzoff and Rechele Brook's (2006) account of gaze following relies less on innate mechanisms, and instead promotes a "starting-state nativism" view, where evolution provides newborns with "discovery procedures" for developing adult common-sense psychology through learning. According to Meltzoff and Brooks, infants learn gaze following by using first-person experiences to, first, distinguish people (material objects that behave as they do) from other things; second, to realize the equivalence between themselves and others (the "Like Me" hypothesis' (Meltzoff, 2005, 2007)); and third, to transpose their experiences with sight to others. Infants thus follow gaze because they realize the equivalence between them and others, and want to see what the other is seeing. Evidence in favor of this view comes from an experiment where infants followed a blindfolded person's gaze less only after experiencing the lack of vision from being blindfolded themselves (Meltzoff and Brooks, 2004).

II.A.5 Bruner - Gaze following as a foundation for theory-of-mind

Jerome Bruner (1995) also proposed a mix of nativistic and learning explanations. He focused on how initial endowments can, through interactions with a social environment, become foundational for (that is, not merely precursors of) subsequent skills. In particular, he saw joint attention as an intermediate point in the developmental path towards a full-fledged theory-of-mind. The initial point in this path consists of two primitives developed within the first few months of life. The first primitive is a construal of people as agents: the understanding of human actions as dedicated to attaining ends. The second primitive is the understanding that arbitrary signs can "stand for" things in the world of experience. When the infant is able to combine these two notions to understand that gestures (or words) can stand for objects in the world, joint attention can be achieved. In the case of joint visual attention, the face is the "gesture" that stands for the objects within the room. Bruner stressed the fact that adults facilitate this learning by standardizing occasions where joint attention happens, as in book reading, mealtime, and greeting/farewell situations.

II.A.6 Tomasello's intentional agents

Michael Tomasello's understanding of joint attention (Tomasello, Kruger, and Ratner, 1993a; Tomasello, 1995) is also based on its foundational role for the development of theory-of-mind abilities. Two phases in this development are proposed: The first, pertaining to joint attention, happens at around one year of age. At this age, infants understand others as intentional agents (that is, their behavior is understood as intentional, and they are understood as having goals and making active choices among behavioral means for attaining those goals), and their perception is understood as attentional (i.e. paying attention to things that will help them attain those goals (Gibson and Rader, 1979)). The second phase happens at around 4 years of life, when infants understand others as mental agents, that is, as having thoughts and beliefs that may differ from reality. While the first phase is considered a cornerstone of social understanding, a 'social-cognitive revolution', the second is seen as simply 'icing on the cake' of social understanding (Tomasello and Rakoczy, 2003). Joint attention is seen as a manifestation of the first phase, and theory-of-mind a manifestation of the second phase.

Tomasello acknowledges that the mechanisms underlying joint attention are largely unknown. But he offers his account of its development in terms of the infant's understanding of its intentionality and the intentionality of others during interactions (Tomasello, 1995): During the first 9 months of age infants learn the distinction between themselves and others and also learn to identify themselves with others, by imitating and reciprocating their behaviors. This can be either innate (Meltzoff and Gopnik, 1993; Gopnik and Meltzoff, 1993) or learned (Moore and Corkum, 1994; Barresi and Moore, 1996). At around 9 months of age, they learn to produce behaviors where means and ends are clearly separated (i.e. they have goals and employ actions for achieving them, such as pulling a cloth to bring closer a toy that rests on top of it (Frye, 1991), or removing the lid from a container to access its contents). In other words, they behave intentionally, through goals. Again, this can be either innate (Trevarthen, 1979, 1993) or learned (Kaye, 1982). With these two abilities, they can understand the intentional activities of other persons based on their new way of interacting with the world intentionally. Gaze following thus would result as the infant tries to look at what the other is looking. Tomasello stresses the power of a structured cultural environment in bringing out the integration of these two abilities, as in the case of chimpanzees raised in humanlike cultural environments that develop joint attention and imitative learning skills (Carpenter et al., 1995; Tomasello et al., 1993b).

II.A.7 Corkum & Moore's learning account

Moore and Corkum (1995) proposed a learning account of gaze following, based on instrumental conditioning: During normal infant-caregiver interaction, the infant is likely to look at the caregiver, and at the same time the caregiver could be looking at some object. The infant becomes distracted and, without necessarily knowing the relation between the caregiver's gaze direction and the location of interesting objects in the room, happens to look at the same object as the caregiver. This experience reinforces the infant's future head turns in the same direction under the same conditions. Social "games" (Trevarthen and Hubley, 1978; Watson, 1972) are likely to boost this learning. In these "games", the mother draws the infant's attention to a salient object such as a toy, and in the process repeatedly turns her head orientation from infant to the object and back.

This account of gaze following was presented as an alternative to the "common sense" view of gaze following, which assumes that the infant understands the psychological relation between the adult and the visual target, and that this understanding of the adult's relation to an object is what makes the infant follow gaze (Baron-Cohen, 1991; Bretherton, 1991; Bruner, 1983; Reddy, 1991; Tomasello, Kruger, and Ratner, 1993a)). Moore and Corkum argue that the common sense view assumes and understanding of the equivalence of self and others that develops only later, during the preschool period. They cite Perner's work (1991), which states that the behavior of infants is governed by a single representation or model of reality, and only after infancy do they develop the capacity for multiple modeling. In particular, a single model does not allow infants to consider simultaneously the orientation of both self and other to some object. In support of their learning account, Moore and Corkum carried out experiments showing that infants who are not spontaneously engaging in joint attention can be conditioned to do so (Corkum and Moore, 1995, 1998). But even while arguing for a learning account of gaze following, they acknowledged that a basic form of joint attention abilities could have been evolved and subsequently refined.

Moore (1999) later gave an account based on the infant's three stages of their control of visual attention: At first, gaze following occurs only if there is a target within the infant's field of view, because the adult's head turn produces an obligatory or reflexive shift in visual orienting in the same direction (Hood, 1995; Johnson, Posner, and Rothbart, 1994). This biases the choice between the two targets to the one that the caregiver is looking at. The second stage happens at around the end of the first year, when infants follow gaze to targets that are out of sight. At this stage, endogenous influences (expectations of finding visually rewarding objects, rather than exogenous influences, which would be the visual saliencies of objects within the field of view) are sufficient to make the infant turn back. In other words, the infant turns back because it is expecting to find something there. In the third stage of gaze following, infants start to use another's eye direction to follow gaze. Moore gives no explanation of the mechanisms at this stage. He simply states that it is likely that turns based on eye direction happen only after turning based on head direction, and that this is probably because infants attribute 'meaning' to eyes at this age.

II.B Computational and Robotic Models of Gaze Following

Computational and robotic models complement the theoretical models described above. The benefit of this type of models is that they ground the notions that are described in theoretical accounts, which are mostly presented verbally. For nativistic accounts, this helps refine their description of how the behavior is achieved, avoiding any unreasonable assumptions. For learning accounts, these models constitute a “working proof” that learning can account for the development of the behavior. They also help explain the progression in the acquisition of the skill, and what might cause it to develop in a different way, or not to develop at all.

Computational simulations of joint attention are attractive with respect to robotic models because they reduce the complexities and costs associated with building robots, although they run the risk of oversimplifying the phenomena of interest. The pros and cons of such models are discussed in more detail in chapter VII.

II.B.1 COG and Infanoid - Pre-programmed joint attention

The COG (Brooks et al., 1999) and Infanoid (Kozima and Yano, 2001; Kozima, 2002) robots (see Fig. II.1) use color saliency, motion, and skin color detectors to detect faces, and eyes within the faces. The eyes’ direction angle is extracted and extrapolated to follow gaze to distal objects, and then motor routines are used for alternating gaze between the object and the face.

These robots represent nativistic versions of gaze following, because the underlying behaviors are pre-programmed, so that the robots exhibit gaze following when first run. As such, they do not give an account of why gaze following develops, and what might cause individual differences in development. It should be noted that the intention in building these robots was to bootstrap developmental processes where the robot communicates with humans in their environment, and they can be better appreciated in their greater role in implementing more complex social understanding skills such as social referencing and theory-of-mind (see Thomaz, Berlin, and Breazeal (2005) and Scassellati (2002)).

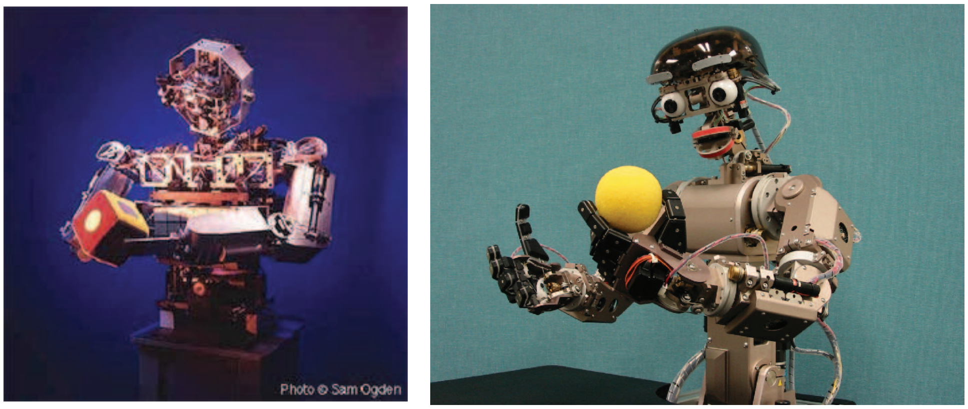


Figure II.1: COG robot (left), and Infanoid robot (right).

II.B.2 AIBO - Joint attention in four-legged robots

AIBO (Artificial Intelligence roBOt) (Fujita and Kitano, 1998) is a dog-like social robot manufactured by Sony, and used to explore learning through interactions with humans. AIBO uses four-legged locomotion, a camera for visual input, two microphones and body sensors. AIBO's implementation of point following is considere here because, even if it is not an instance of *gaze* following, the two behaviors are very similar.

To implement joint attention, two AIBO robots were placed facing each other, with one of them pointing at an object (Hafner and Kaplan, 2005). The other robot's visual input was analyzed: the left and right sides of the resulting image were processed using four different filters (two brightness threshold filters and an edge detection filter), resulting in four different images per side. Three different operators were applied to each image (two center of mass operators, and another consisting of a summation of values), resulting in 12 features per side. These were subtracted from each other to get a resulting 12-valued vector. Three values of the vector were selected according to their capacity to differentiate between pointing to the left or to the right, using a three-layer perceptron trained using supervised learning (Duda, Hart, and Stork, 2001). The robot could distinguish if the other robot was pointing to the right or to the left.

Kaplan's implementation is of value because it is an instance of learned joint attention. However, the resulting behavior is very simple, consisting of a turn to either the right or the left. It does not reach the complexity of the development of joint attention in infants. Additionally, this implementation of joint attention behavior relies on a signal, given by a human, correcting the robot each time it follows gaze incorrectly (i.e. it is based on supervised learning). This contrasts with human infants, who learn to do so without such signals.

II.B.3 Neural Systems Group - Probabilistic imitation learning

The Neural Systems Group at the University of Washington implemented a gaze following robot (Hoffman, Grimes, Shon, and Rao, 2006) as part of a larger effort to program imitation based on Meltzoff and Moore's Active Intermodal Mapping (AIM) hypothesis (Meltzoff and Moore, 1977), where imitation is seen as a goal-directed matching process by which infants compare their motor states with states observed in the adult's behavior. The robot implements gaze following using a Bayesian framework (Jordan, 1998): Instructor-based cues (saccadic eye movements, hand gesture direction, head gaze direction, etc) are represented using variables $A_1 ... A_n$; object properties are represented using variables $O_1 ... O_n$; the caregiver's true focus of attention, which cannot be directly observed, is represented using variable X (which specifies, for example, a discrete object identifier); the instructor's preferences with respect to saliency components (size, color, brightness, etc) are represented using variable S. Different instructors have different preferences, and these are estimated through interactions with the caregiver using the Expectation Maximization algorithm (Dempster, Laird, and

Rubin, 1977). Fig. II.2 shows the relationships between variables. The probability that the instructor is looking at a certain location is thus defined as:

$$P(X|A_{1\ldots n},\bar{O}_{1\ldots k},S) = P(X|S,\bar{O}_{1\ldots k})P(A_1|X)...P(A_n|X)$$

where $\bar{O}$ is the value of O inferred from the scene I by extracting saliency information. To test the model, an instructor faces the robot, with a table between them and objects placed on the table. A Biclops active stereo vision head is used to gather camera images and detect faces (Wu et al., 2000). Once detected, the instructor's face is tracked using a feature-based object detection framework, keeping it within a bounding box using the Meanshift algorithm (Comaniciu et al., 2000) and Kalman filters (Kalman, 1960). Then, a likelihood over head and tilt angles is transformed to egocentric coordinates. The resulting estimated direction (A) is calculated from the visual appearances of the objects (values of $\bar{O}$) as well as the robot's estimation of the instructor's preference (S), and used to decide on the robot's action. This results in joint attention after learning takes place.

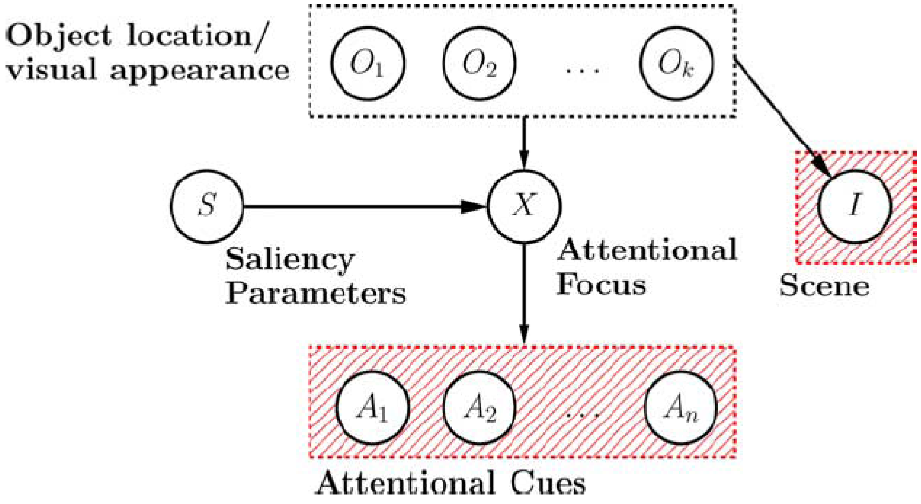


Figure II.2: Bayesian framework for gaze following designed by the Neural Systems Group.

This approach to modeling gaze following is valuable because it is well grounded within a Bayesian framework. However, the initial gaze turn by the robot is not learned, but is pre-programmed. And only the experimenter's head direction, and not the object saliencies, are considered for the robot's initial gaze shift, making all the objects *de facto* outside the robot's field of view. This makes it impossible to replicate experiments where the object may be outside the infant's field of view, such as the ones in Butterworth and Cochran (1980) and Butterworth and Jarrett (1991).

II.B.4 Nagai - Cognitive developmental robotics

Yukie Nagai (Nagai et al., 2003, 2006) bases her work on Minoru Asada's cognitive developmental robotics (CDR) (Asada et al., 2001), a constructivist approach to building robots. CDR advocates building robots to help gain an insight into how cognitive abilities emerge from the interaction of a physical robot with its environment and other people. This contrasts with the more common approach of building robots based on the designer's understanding of how to solve the relevant task. By placing robots from the start in the kind of environments that they will work on while solving real problems, these robots arguably have fewer problems scaling their abilities to solve real-world tasks, a common problem in traditional robot building.

Nagai applied the CDR paradigm to program robots to achieve joint attention through "self-learning". A human caregiver is positioned in front of the robot, and several salient objects are positioned in the room. A *salient feature detector* (refer to Fig. II.3) takes the robot's visual input and extracts color, edge, and motion components, and detects human faces. These components drive a *visual feedback controller*, which outputs a displacement vector that the motor control uses to redirect the robot's camera to the most salient object based on these features. This motor control command competes with a similar one generated by the *learning with self-evaluation module*, consisting of a *learning module* and an *internal evaluator*. The learning module is a three-layer neural network (Bishop,

1995) which takes as input the caregiver's face image, detected by the salience feature detector, and the angle of the camera head. The neural network's output is a displacement vector. A *gate* takes the displacement vectors from the visual feedback controller and the learning module, and chooses one probabilistically, with an initial bias towards using the motor control command based on the salient features, and a later bias towards using the motor control command based on the learning module. The internal evaluator evaluates the action as a success if an object is located in the center of the screen after a turn. This is used as a signal to train the neural network through backpropagation. Although in many cases the robot will "succeed" while looking at an object not looked at by the caregiver, this does not prevent the robot from learning to follow gaze.

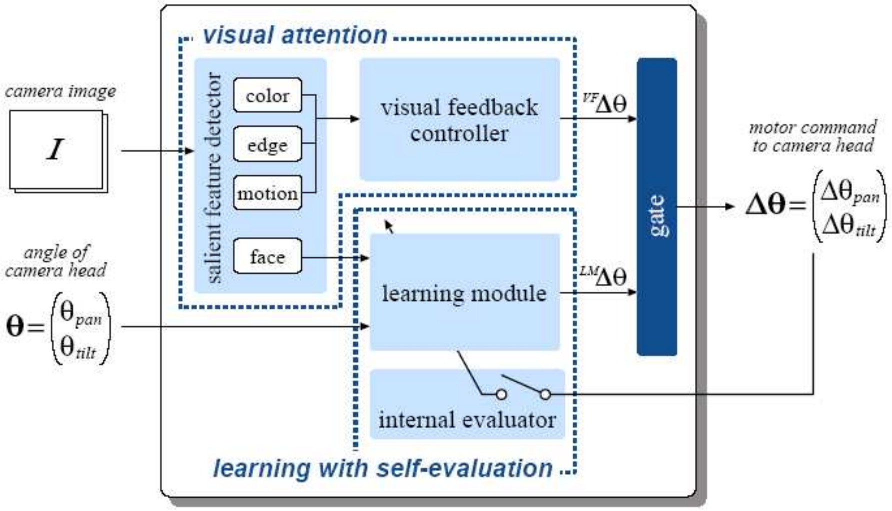

visual attention
camera image
I
salient feature detector
color
edge
motion
face
visual feedback controller
$^{VF}\Delta\theta$
angle of camera head
$\theta = \begin{pmatrix}\theta_{pan} \\ \theta_{tilt}\end{pmatrix}$
learning module
$^{LM}\Delta\theta$
internal evaluator
gate
motor command to camera head
$\Delta\theta = \begin{pmatrix}\Delta\theta_{pan} \\ \Delta\theta_{tilt}\end{pmatrix}$
learning with self-evaluation

Figure II.3: Details of the constructive model of joint attention developed by Yukie Nagai (Nagai et al., 2003, 2006).

The robot develops gaze following through three stages (refer to Fig. II.4), which are meant to resemble Butterworth's *ecological*, *geometric*, and *representational* stages (Butterworth and Cochran, 1980; Butterworth and Jarrett, 1991) (these stages are described in more detail in chapter IV). In the first stage (top row), the gate chooses mainly $^{VF}\Delta\theta$ (the output of the visual attention module, derived from the visual features) over ${}^{LM}\Delta\theta$ (the output from the learning module). Although Nagai describes this as equivalent to Butterworth's ecological stage, in Butterworth's experiments the infant did follow the experimenter's eye direction. This stage, therefore, would better match the infant's behavior before the ecological stage. In the second stage (middle row), ${}^{VF}\Delta\theta$ and ${}^{LM}\Delta\theta$ are chosen with non-zero probabilities. The robot has had a chance to train the neural network and therefore follow gaze. In the cases where there are two objects in the field of view, the robot will sometimes choose to act based on ${}^{LM}\Delta\theta$, achieving gaze following. In the case where the object is outside the field of view, and the robot chooses to act based on ${}^{LM}\Delta\theta$, then correct gaze following will only happen when no distracter objects appear after the robot makes the initial turn. Otherwise, the saliency of the distracter might cause the robot to miss the object looked at by the caregiver. Nagai considers this equivalent to the geometric stage, where gaze is followed correctly as long as the object is within the infant's field of view. However, in the case presented by Nagai, these two objects were outside the infant's field of view, making it different to the setting used by Butterworth. In the third stage (bottom row), since the robot mostly chooses ${}^{LM}\Delta\theta$, gaze can be followed to objects that are outside the field of view, because the distracter's saliency will have no effect, as ${}^{VF}\Delta\theta$ is almost never chosen. This corresponds to Butterworth's *representational* stage, with the slight difference that distracter objects within the infant's field view do stop the infant from following gaze.

Nagai's robotic model is an important step in the construction of models of gaze following, especially since it learns to follow gaze without the need for a supervision signal telling it when it has achieved the goal and when not. However, although the stages that the robot goes through represent a developmental shift from following saliencies to following gaze, Butterworth's experimental setup and results are not exactly replicated. The robot also has not been used to explore other central aspects of gaze following such as the relative importance of eye and head direction cues, as described in Corkum and Moore (1995) and Meltzoff and

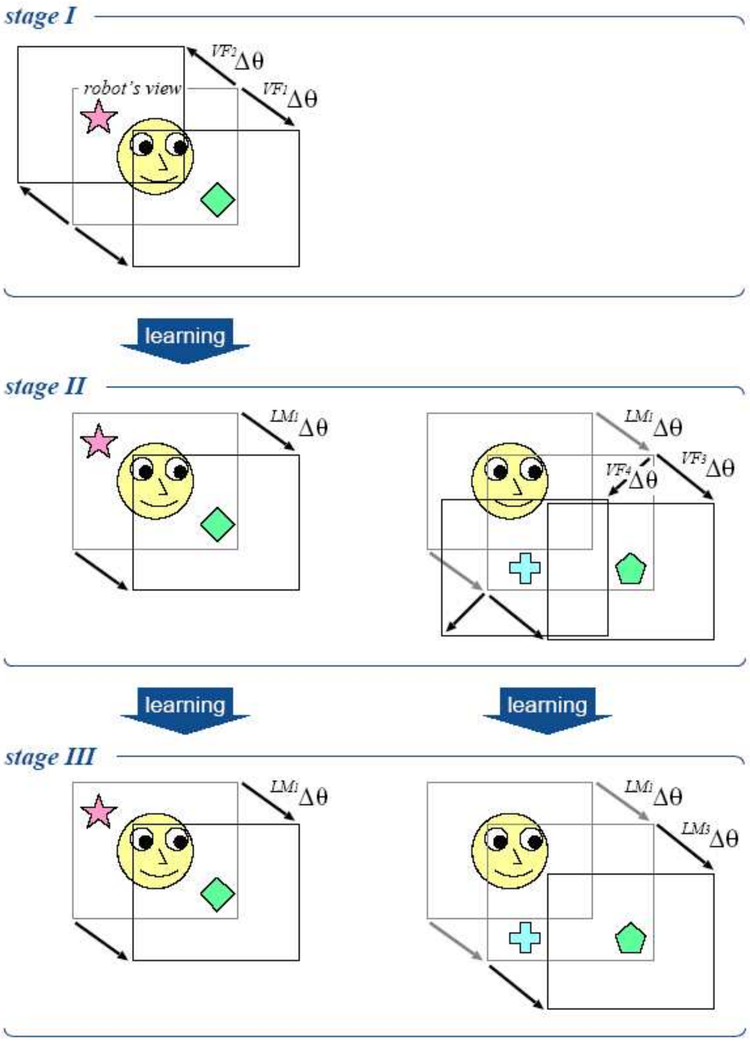


Figure II.4: The staged learning process of Nagai's robot's joint attention. See text for description of stages.

Brooks (2006).

II.B.5 MESA project - The Basic Set hypothesis

The MESA (Modeling the Emergence of Shared Attention) project at the University of California, San Diego (UCSD) focuses on the study of joint attention. It is based on a set of premises named the Basic Set Hypothesis (Fasel, Deák, Triesch, and Movellan, 2002), which states that the following elements are sufficient (although not necessary) for joint attention to happen:

- A set of *motivational biases*, in particular a preference for social stimuli such as human faces (Dannemiller and Stevens, 1988).
- A *structured environment* providing strong correlation between where parents look and where interesting things are.
- *Habituation* (Stanley, 1976) as a basic learning mechanism, and as the mechanism by which the infant goes from early dyadic interactions with the caregiver to triadic interactions involving the objects around them.
- A *learning mechanism* such as temporal difference learning (Sutton and Barto, 1998), to learn the temporal structure of predictable, contingent interactions between infant and caregiver. Temporal difference learning has been proposed within the computational neuroscience community as a model of learning in the brain (Houk, Davis, and Beiser, 1995; Dayan, Kakade, and Montague, 2000; Schultz, Dayan, and Montague, 1997; Doya, 2000).

The Basic Set Hypothesis is based on a dynamical systems approach to the modeling of cognitive phenomena (Thelen and Smith, 1994), which contrasts with theories of innate mechanisms such as those advocating Fodorian modules. Dynamical systems do not start with complex representations, but instead rely on the environment to form them, either through the environment's dynamics, or internally as direct responses to inputs.

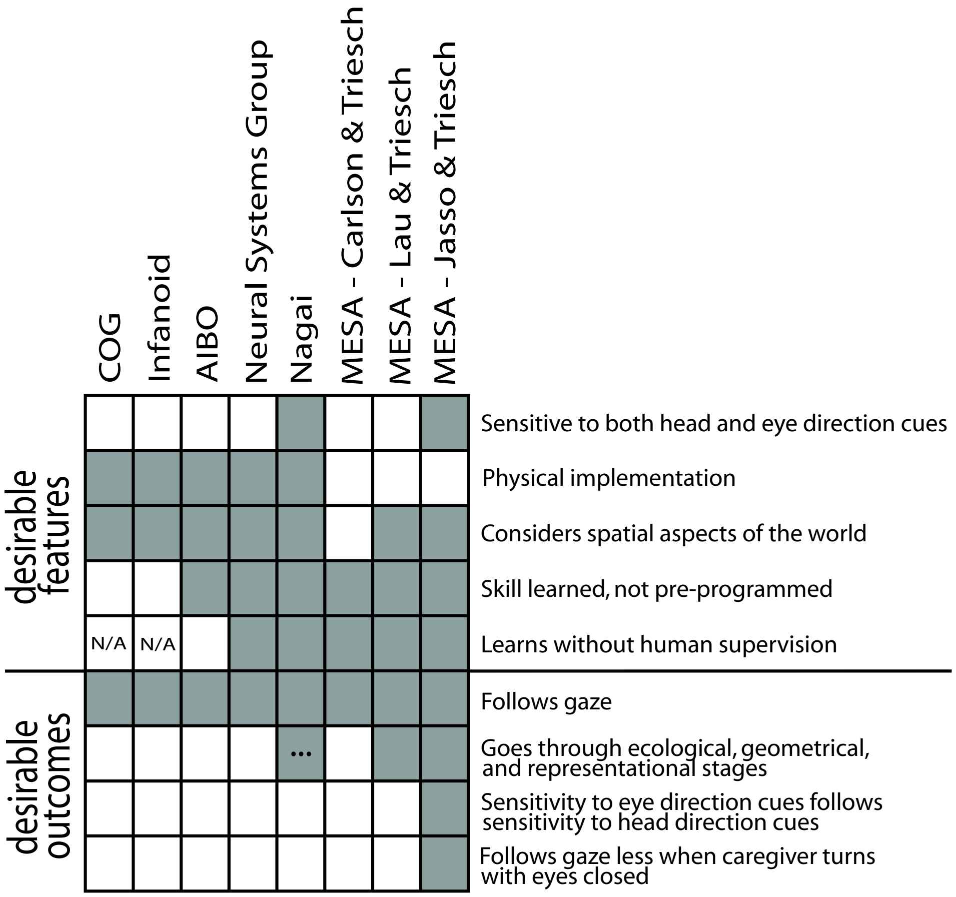
The first gaze following model built within the MESA project was created by Jochen Triesch, Eric Carlson, and Christof Teuscher (Carlson and Triesch, 2003; Triesch, Teuscher, Deák, and Carlson, 2006b). The model showed that gaze following behaviors can emerge given the Basic Set. The environment was modeled as a set of discrete regions occupied by the infant, the caregiver, and objects. With time, the model learned to first look at the caregiver, and then, when it habituates to the caregiver, to follow its gaze to the appropriate region in space where the object was located, as indexed by the caregiver's gaze direction.

The second model, created by Boris Lau and Jochen Triesch (Lau and Triesch, 2004) is similar, but with a spatial representation of the environment, using a body-centered coordinate system. A Hebbian-like learning rule is used to strengthen the connections between visual inputs and the locations where visual saliency is encountered as a result of actions. The model follows Butterworth's three stages (Butterworth and Cochran, 1980; Butterworth and Jarrett, 1991): It first learns to follow gaze to objects within its field of view and then to objects outside its field of view. It also learns to differentiate between the target and other objects that are on the correct side of the room but not on the caregiver's line of sight.

In the first MESA model, regions are discrete and hold no spatial relationship among each other, so the model cannot be used to replicate spatial aspects of Butterworth's experiments. The second model has not been used to replicate more complex aspects of gaze following, such as those having to do with conflicting eye and head direction cues (Corkum and Moore, 1995; Meltzoff and Brooks, 2006).

II.C Discussion

Fig. II.5 lists desirable features and outcomes for robotic and computational models, as described next. The Jasso & Triesch model, presented in this dissertation, is discussed in the next chapters and is omitted from this discussion.



	COG	Infanoid	AIBO	Neural Systems Group	Nagai	MESA - Carlson & Triesch	MESA - Lau & Triesch	MESA - Jasso & Triesch	
desirable features					▨			▨	Sensitive to both head and eye direction cues
	▨	▨	▨	▨	▨				Physical implementation
	▨	▨	▨	▨	▨		▨	▨	Considers spatial aspects of the world
			▨	▨	▨	▨	▨	▨	Skill learned, not pre-programmed
	N/A	N/A		▨	▨	▨	▨	▨	Learns without human supervision
desirable outcomes	▨	▨	▨	▨	▨	▨	▨	▨	Follows gaze
					▨ ...		▨	▨	Goes through ecological, geometrical, and representational stages
								▨	Sensitivity to eye direction cues follows sensitivity to head direction cues
								▨	Follows gaze less when caregiver turns with eyes closed

Figure II.5: List of robotic and computer models of gaze following, and desirable features and outcomes. The Jasso & Triesch model is described in the next chapter.

II.C.1 Desirable features

- *Sensitive to both head and eye direction cues*: Infants are sensitive to head and eye direction cues (Corkum and Moore, 1995; Meltzoff and Brooks, 2006). This should be included in gaze following models to explain how these cues get incorporated into the behavior with age. The COG and Infanoid robots use eye direction and not head direction, and Neural Systems Group robot uses head direction and not eye direction. Nagai's robot uses a complete image of the experiment's face as input, and therefore constitutes a combination of eye and head direction cues. The two MESA models described in this chapter do not distinguish between the two cues, and only talk about the caregiver's gaze direction.

- *Physical implementation*: Physical (robotic) models are preferred because they make almost no simplifications about how they are to work in real life environments like the ones where infant-caregiver interactions happen. (However, see pros and cons of using robotic systems, in chapter VII.)

- *Considers spatial aspects of the world*: To replicate the experiments that Butterworth carried out (see subsection II.A.1), the model should be able to distinguish between objects placed at different distances and at different angles, and should have a limited field of view. All robotic systems reviewed have this (but, see note on the Neural Systems Group robot, above) as well as the computational model proposed by Boris Lau and Jochen Triesch. The model proposed by Carlson and Triesch only represents space as a categorical value, and cannot be used to replicate Butterworth's experiments.

- *Skill learned, not pre-programmed*: As described in chapter I, it is desirable to have the model learn to follow gaze instead of having the skill explicitly pre-programmed. All models reviewed incorporated some form of learning, except the COG and Infanoid robots.

- *Learns without human supervision*: It is desirable that learning models do not require human supervision, because it makes training them easier, and because human infants learn this way. AIBO is the only learning model requiring human supervision. The COG and Infanoid robots do not learn at all.

II.C.2 Desirable outcomes

- *Follows gaze*: This is the most basic outcome. All models follow gaze.
- *Goes through ecological, geometric, and representational stages*: A model that goes through these stages (as described in subsection II.A.1 and in more detail in chapter IV) helps explain the development of spatial components of gaze following as used for referential purposes. Nagai's model has been shown to implement these stages, although, as described in section II.B.4, the mapping to Butteworth's stages was not rigorous. Lau & Triesch's model goes through these stages.
- *Sensitivity to eye direction cues follows sensitivity to head direction cues*: To properly model infant gaze following development, systems should first exhibit gaze following based on the caregiver's head direction, and only gradually should the eye direction be incorporated into its behavior (Corkum and Moore, 1995, 1998) (as described in detail in chapter V). Nagai's model could be used to test this transition, since it is sensitive to the eye and head direction cues. However, since the two cues are not explicitly separated, it does not seem likely that it would replicate this transition as it is.
- *Follows gaze less when caregiver turns with eyes closed*: Related to the previous point, infants transition to following gaze less when the caregiver turns with her eyes closed (Meltzoff and Brooks, 2006). Again, Nagai's model might be able to do so. But none of the models attempt to replicate this transition.

III

A Reinforcement Learning Model of Gaze Following

This chapter introduces a new computational model of gaze following. The model takes elements from the two MESA models described in the previous chapter: Carlson & Triesch's model (Carlson and Triesch, 2003; Triesch, Teuscher, Deák, and Carlson, 2006b) and Lau & Triesch's model (2004). Like them, it is based on the MESA project's Basic Set Hypothesis (Fasel, Deák, Triesch, and Movellan, 2002). From the first it incorporates a biologically plausible reinforcement learning algorithm (Sutton and Barto, 1998) that has proven to be useful for exploring aspects of the development of gaze following (Carlson and Triesch, 2003; Triesch, Teuscher, Deák, and Carlson, 2006b; Teuscher and Triesch, in press). From the second it incorporates a spatial representation of the environment, necessary to replicate many gaze following experiments. The value of the model lies in its ability to replicate the key experimental results used to measure gaze following, as described in the next chapters. For this reason, it can be used to explore issues about learning vs. nativistic accounts and the understanding of other's attention during gaze following, as discussed in chapter I.

First, the model is described in detail, showing how basic gaze following performance is achieved by replicating Scaife & Bruner's original experiment (1975). Then, an experiment is set up to show how the model transitions from looking mostly at salient objects to also following gaze.

III.A Proposed Model of Gaze Following

III.A.1 Modeling the environment

The environment is modeled as follows: Infant and experimenter are positioned facing each other with a 40 cm separation between them (see Fig. III.1). Objects can be placed anywhere except on the same location as the infant or caregiver. Time is discretized into steps of 1 second.

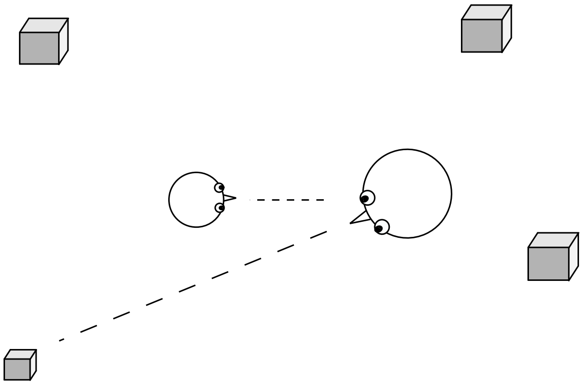


Figure III.1: Modeling the environment: Infant and caregiver sit facing each other, with objects placed around them.

The following variables are calculated every time step from the state of the environment, and used for the different calculations described below (refer to Fig. III.2):

- $\varphi_I \in [0°,\ 360°]$ is the infant's heading ($\varphi_I = 0°$ corresponds to the infant looking at the caregiver).
- $\varphi_H \in [0°, 360°]$ is the caregiver's head direction ($\varphi_H = 180°$ corresponds to the caregiver' head directed towards the infant).
- $\varphi_E \in [0°, 360°]$ is the caregiver's eyes direction ($\varphi_E = 180°$ corresponds to the caregiver looking at the infant).
- $\varphi_{o_i} \in [0°, 360°]$ is the angle of object i from the infant's point of view (i.e. the value of φ_I corresponding to the infant looking at object o_i).

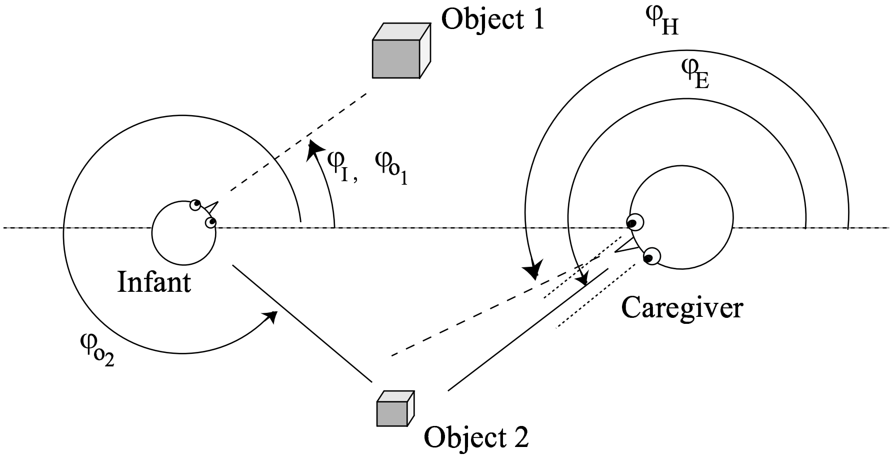


Figure III.2: Variables extracted from the state of the environment: φ_I (infant's heading), φ_H (caregiver's head direction), φ_E (caregiver's eyes direction), and φ_{o_i} (object o_i's angle from the infant's point of view).

Φ_{o_i} is the visual saliency of object o_i. Φ_C is the caregiver's visual saliency when facing the infant. This saliency is set to half when the caregiver is not looking at the infant, modeling infants' preference for looking at gaze directed at them than diverted elsewhere (Farroni, Csibra, Simion, and Johnson, 2002). Φ_I is the infant's visual saliency. The infant's field of view, $FOV \in [0, 360^o]$, specifies the extent of the visible area with respect to the infant's gaze direction.

III.A.2 Infant visual system

The infant's visual input is processed by three different systems (see Fig. III.3 left): a saliency map ($\mathbf{s}$), a head direction detector ($\mathbf{h}$), and an eyes direction detector ($\mathbf{e}$). These are described next:

Saliency Map ($\mathbf{s} = [s_1, ..., s_{96}]$) Indicates the presence of visual saliency in a body-centered coordinate system with 96 different regions in space, along 24 heading ranges and 4 depth ranges. Heading 1 corresponds to heading angles between -7.5° and 7.5°, heading 2 corresponds to angles between 7.5° and 22.5°,

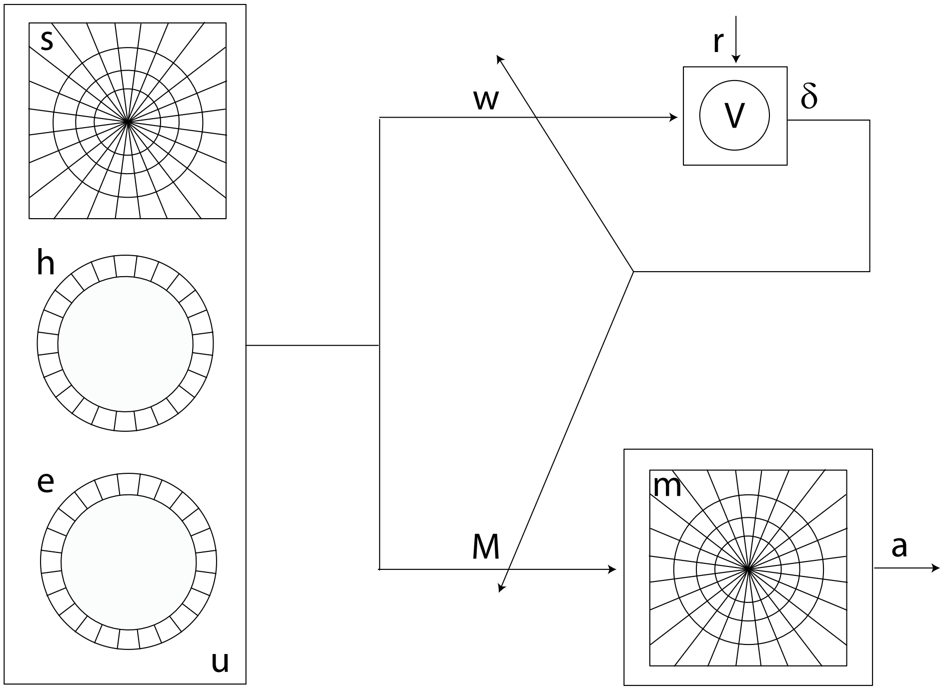


Figure III.3: Details of the infant visual system (left) and of the actor-critic reinforcement learning model (right). Features calculated from the *Saliency Map* **s**, *Caregiver Head Direction* **h**, and *Caregiver Eyes Direction* **e** are combined into **u**, weighted using **w** and added into V to calculate the value of the present state. They are also weighted using **M**, added into **m**, and passed through a softmax selection formula to calculate the next action a.

and so on, covering all 24 different headings. Depth 1 corresponds to distances (from the infant's perspective) of up to 0.8 meters away, depth 2 corresponds to distances of 0.8 to 1.2 meters, depth 3 corresponds to distances of 1.2 to 1.7 meters, and depth 4 corresponds to distances of more than 1.7 meters.

The saliencies of objects and caregiver within the infant's field of view ($\varphi_I - FOV/2 \leq \varphi_{o_j} \leq \varphi_I + FOV/2$) are added to the element in $\mathbf{s}$ corresponding to their location (heading and depth), after foveation and habituation are calculated:

Foveation causes an object's perceived saliency to decay as it falls outside the infant's center of vision according to the following formula. The formula used is based on the contrast sensitivity function proposed by Daly et al. (1999):

$$S(x, y) = \frac{1}{1 + k_{Ecc} \cdot \theta_E(x, y)} \tag{III.1}$$

where S is the visual sensitivity of an image position of an object (x, y), θ_{Ecc} is the eccentricity in visual angle of the object, and k_{Ecc} is a constant that defines how the sensitivity diminishes with eccentricity. This formula captures reduction of bandwidth and peak sensitivity as a function of eccentricity (Virsu and Rovamo, 1979). k_{ECC} is set to 0.24 based on a fit on data sets from Virsu and Rovamo (1979) and Johnston (1987).

The formula used here is a modification of the above, scaled with an offset of 0.2:

$$foveation(\theta) = 0.2 + 0.8\frac{1}{1 + k_{Ecc} \cdot \theta} \tag{III.2}$$

where $\theta \in [0, FOV/2]$ is the eccentricity in visual angle of the object. The offset prevents values from decaying to close to zero when objects are in peripheral vision (i.e. "in the corner of the eye"), which helps replicate some of the gaze following experimental results where a distracter object at the periphery of vision captures the attention of the infant. The resulting foveation is depicted in Figure III.4.

The infant habituates separately to each object, according to the discretized version of the following exponential decay formula proposed by Stanley

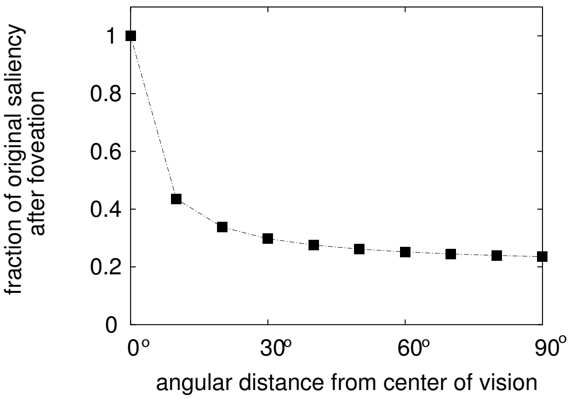


Figure III.4: Graphical depiction of the foveation formula.

(1976):

$$\tau_H \frac{d\phi_{o_j}(t)}{dt} = \alpha_H(\Phi_{o_j} - \phi_{o_j}(t)) - S_{o_j}(t) \qquad \text{(III.3)}$$

where $\phi_{o_j}(t)$ is object j's habituated saliency at time t and Φ_{o_j} its original, dishabituated, saliency; $S_{o_j}(t)$ is equal to Φ_{o_j} if the infant is looking at object j at time t and 0 otherwise; τ_H is a time constant that specifies the rate of habituation (a smaller τ_H resulting in faster habituation); and α_H controls the level of long-term habituation. A similar formula applies for ϕ_C and ϕ_I, the habituated saliencies of the caregiver (Φ_C) and the infant (Φ_I), respectively.

Finally, when an element s_i of $\mathbf{s}$ is outside the infant's field of view, its new value is calculated by multiplying the previous value by a constant d ($0 < d < 1$), a "memory decay" factor. This enables the model to temporarily remember recently observed states of the world.

The top section of Fig. III.5 shows an example setting and the resulting value of s.

The exact formula for calculating $\mathbf{s}$ is: $s_i = S_O + S_C + S_{M_i}$; where

- $S_O = \sum_{j=1}^{N} S_{o_j}$ and $S_{o_j} = \phi_{o_j} foveation(\theta_{o_j})$ if o_j is within the infant's field of view, 0 otherwise, $\theta_{o_j} = |\varphi_I - \varphi_{o_j}|$ being the angular distance of the object from the center of vision,

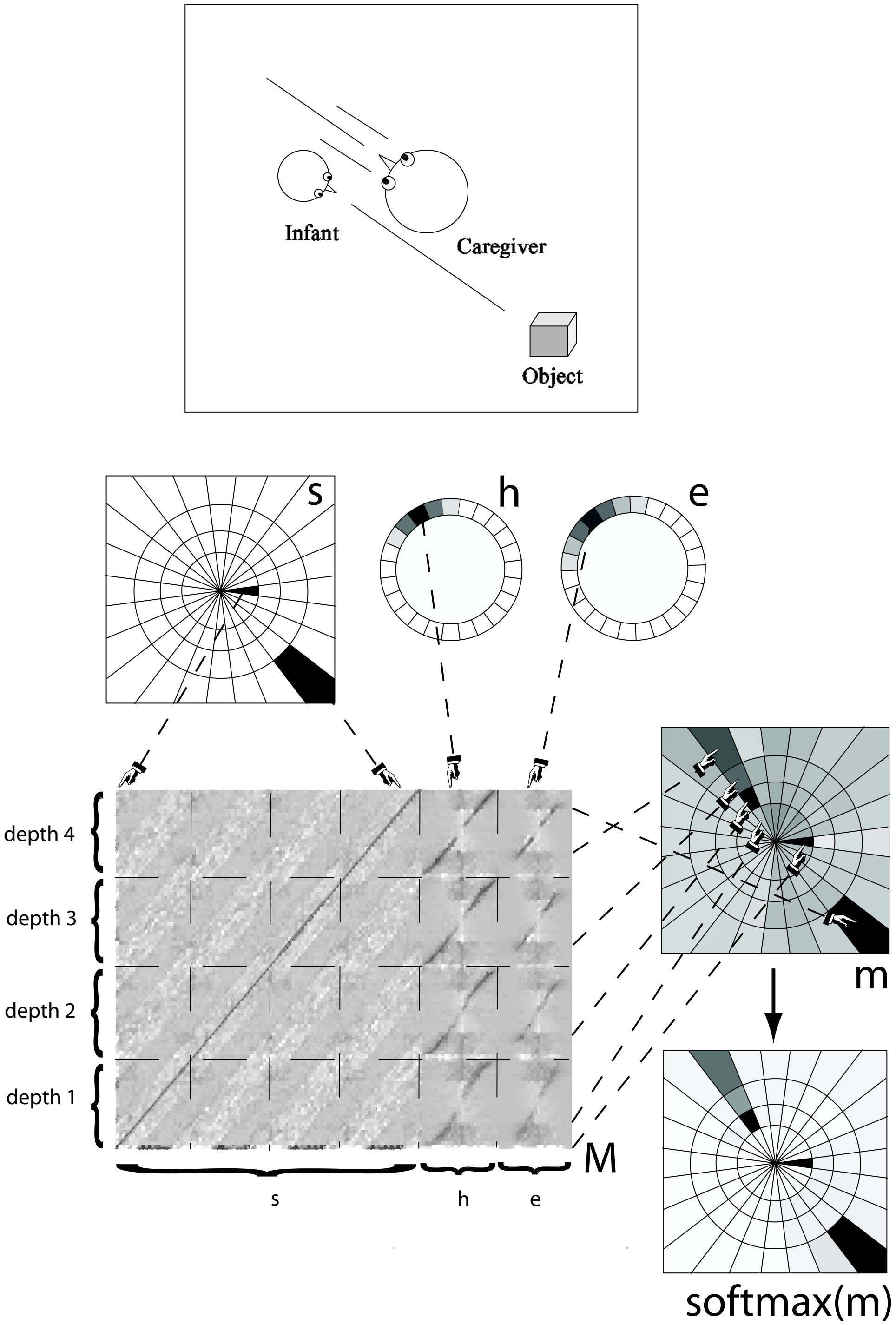


Figure III.5: Selecting an action: The values for the infant's visual input (**s**, **h**, and **e**) are multiplied by the weight matrix **M** (darker sections of **M** correspond to higher values) to get the value of **m**. A softmax selection on **m** is used to calculate the probabilities of choosing different actions. Values in **h** and **e** result from traces of memory, from the infant looking at the caregiver in previous time steps.

- $S_C = \phi_C foveation(\varphi_I)$ if the caregiver is present and within the infant's field of view; 0 otherwise,
- $S_{M_i} = s_i(t-1)d$ if the location is outside the infant's field of view, 0 otherwise.

Our assumption of a body-centered representation (in contrast to a retinotopic one) is not physiologically accurate but it frees us from having to model coordinate transformations between different coordinate systems (although it is an interesting question in its own right when and how infants learn to compute certain coordinate transformations).

Head Direction Detector ($\mathbf{h} = [h_1, ..., h_{24}]$) Indicates 24 possible caregiver head directions as perceived by the infant. Heading ranges are similar to those in **s** (heading 1 corresponds to heading angles between -7.5° and 7.5°, heading 2 corresponds to angles between 7.5° and 22.5°, and so on). If the infant is looking at the caregiver, the value of each h_i is calculated according to a exponential decay, so that the closer h_i is to the caregiver's heading (φ_H), the higher its value. **h** is normalized (using linear scaling) so that the sum of all h_i add to 1. While adults are very good at detecting the gaze direction of others (Anstis, Mayhew, and Morley, 1969; Cline, 1967; Gibson and Pick, 1963), being able to detect gaze deviations of just 1.4° at a distance of just over 1 m (Cline, 1967), the development of this ability from infancy has not been systematically studied. Newborns are known to differentiate between direct and averted gaze (Farroni, Csibra, Simion, and Johnson, 2002), and basic gaze following experiments such as Scaife and Bruner's (1975) and Butterworth's (Butterworth and Cochran, 1980; Butterworth and Jarrett, 1991) show a capacity to distinguish between right and left-facing head directions at around 6 months of age. This ability increases with time, so that at 12 months they can discriminate a 25° difference in gaze direction to two objects (von Hofsten et al., 2005). This progression is captured in the model by having the exponential decay be gentler at the beginning of learning, and progressively sharper with time.

If the infant is not looking at the caregiver, then the values of **h** are calculated by multiplying the previous value by the memory constant d, (the same as in the calculation of **s**), to enable the model to temporarily remember recently observed head directions of the caregiver.

The top section of Fig. III.5 shows an example setting and the resulting value of **h**.

The exact formula for calculating **h** is: $h_i = H_C + H_{M_i}$, where

- $H_C = exp(-(\varphi_H - \theta_{I_i})^2/\sigma_H{}^2)$ if the caregiver is present and the infant is looking at the caregiver ($\varphi_I = 0°$), with θ_{I_i} being the angle corresponding to heading i's center ($\theta_{I_1} = 0°$, $\theta_{I_2} = 15°$, $\theta_{I_3} = 30°$, ... $\theta_{I_{24}} = 345°$) and σ_H being a parameter that specifies the exponential decay; 0 otherwise;
- $H_{M_i} = h_i(t-1)d$ if the caregiver is absent or outside the infant's field of view, 0 otherwise.

with a posterior scaling of all h_i so that $\sum_{i=1}^{24} h_i = 1$.

Eyes Direction Detector ($\mathbf{e} = [e_1, ..., e_{24}]$) Similar to **h**, but computed with the caregiver's eye direction (φ_E) instead of head direction (φ_H), and with a different exponential decay parameter (σ_E instead of σ_H). Additionally, when the caregiver is present and within the infant's field of view but turning back ($\varphi_C < 90°$ or $\varphi_C > 270°$), all values e_i are set to zero. This reflects the fact that when the caregiver is facing backwards with respect to the infant, the eyes are not visible.

The top section of Fig. III.5 shows an example setting and the resulting value of **e**.

Such representations of head and eye direction may be found in the superior temporal sulcus (STS) in monkeys, and are likely to exist in humans, too (Jenkins, Beaver, and Calder, 2006). Separate mechanisms for the caregiver's head pose and eye direction allow us to capture the development of the infant's differential sensitivity to these cues.

III.A.3 Reinforcement learning model

The infant's visual system serves as input to an actor-critic reinforcement learning system (Dayan and Abbott, 2001) that drives actions.

The *critic* (see Fig. III.3, upper right) approximates the value of the current state as $V(t) = \mathbf{w}(t)\mathbf{u}(t)$ where $\mathbf{w}(t) = (w_1(t), w_2(t), ..., w_{N_s}(t))$ is a weight vector, $\mathbf{u}(t) = (\mathbf{s}(t), \mathbf{h}(t), \mathbf{e}(t))^T$ is the value of the input features from the visual system at time t, and N_s is the number of features ($N_s = dim\ \mathbf{s} + dim\ \mathbf{h} + dim\ \mathbf{e} = 96 + 24 + 24 = 144$). The weight vector $\mathbf{w}(t)$ is updated according to the formula:

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \eta\delta(t)\mathbf{u}(t), \tag{III.4}$$

where η is the learning rate, and $\delta(t)$ specifies the temporal difference error, defined as the difference between the immediate reward received plus the estimated future discounted reward, minus the current estimated value of the state:

$$\delta(t) = r(t) + \gamma V(t+1) - V(t), \tag{III.5}$$

where $r(t)$ is the reward at time t, $V(t+1)$ the estimated value of the new state after taking the action, and γ the reward discount factor.

The *actor* (see Fig. III.3, lower right) specifies the action to be taken, directing the infant's attention to one of 24 possible different headings and one of four different depths, with a total of $N_a = 96$ different possible actions ($A = (H, D)$, $H \in \{0°, 15°, 30°, ..., 345°\}$, $D \in \{0.4, 1.0, 1.45, 2.0\}$), where A is the action, and H and D are the heading and depth, respectively, where attention is directed to. The action is chosen probabilistically according to the softmax decision rule:

$$P[a] = \frac{\exp(\beta m_a)}{\sum_{a'=1}^{N_a} \exp(\beta m_{a'})}, \tag{III.6}$$

m_a being the action value parameter for action a for the present state: $\mathbf{m} = \mathbf{M}\mathbf{u}$, where $\mathbf{M}$ has as many columns as there are input features and as many rows as there are actions. A higher value of m_a increases the chances of selecting action a.

β is an "inverse temperature" parameter, with a larger value increasing exploitation over exploration. **M** is updated according to:

$$M_{a'b}(t+1) \leftarrow M_{a'b}(t) + \epsilon(\delta_{aa'} - P[a'; \mathbf{u}(t)])\delta(t)u_b(t), \tag{III.7}$$

where η is the same learning rate as above, $\delta(t)$ is the critic's temporal difference error (defined above), a is the action taken, $P[a'; \mathbf{u}(t)]$ is the probability of taking action a' at state $\mathbf{u}$(t), and $\delta_{aa'}$ is the Kronecker delta, defined as 1 if $a = a'$, 0 otherwise.

Reward is obtained as the saliency of the position where attention is directed to after the action is taken and **s** updated with the result of the action (the value of **s** corresponding to the depth/heading of the selected a, but in the next time step, and with a foveation corresponding to the new φ_I). The definition of salience as reward is based on studies of infant visual expectations and the organization of their behavior around these expectations (Haith, Hazen, and Goodman, 1988).

III.A.4 Training and testing scheme

The simulation starts with the infant and caregiver in the middle of the room, 40 cm apart, facing each other. N_o objects are placed in the room, where N_o is drawn from a geometric probability distribution with average $\bar{N}_o$:

$$P(N_o = k) = (1 - 1/\bar{N}_o)^{k-1}(1/\bar{N}_o)$$

Objects are placed randomly around the infant with distances from the infant taken from a radially symmetric normal probability function with standard deviation of σ_o. The saliency Φ_{o_i} for each object i is drawn from an exponential probability with average $\bar{\Phi}_o$:

$$P(\Phi_{o_i} = x) = (1/\bar{\Phi}_o) \cdot exp(-1/\bar{\Phi}_o \cdot x)$$

After a number of time steps drawn from a geometric probability function with average $\bar{T}_{objects}$, all objects in the room are removed and replaced by new objects, with positions and saliencies drawn randomly as described above. Additionally, after a number of time steps drawn from a geometric probability function with average $\bar{T}_{present}$, the caregiver leaves the room. The caregiver returns to continue interacting with the infant after a number of time steps drawn from another geometric probability function with average $\bar{T}_{absent}$.

The simulation is run for *15,000,000* training steps (roughly corresponding to 172 days of a wake infant), during which gaze following develops. During training, the caregiver always looks at the most salient point in the room, which in some cases will be the infant. The caregiver's perceived saliencies are mediated by the same foveation and habituation mechanisms (with identical parameters) as in the infant's visual system. The caregiver's head direction is slightly offset from that of the eyes according to a Gaussian distribution with $\sigma = 5°$ and $\mu = 0°$. This offset is recalculated for every gaze shift that the caregiver does. This reflects the fact that eyes and head are not always perfectly aligned, and corresponds to values observed in naturalistic settings (Hayhoe, Land, and Shrivastava, 1999). The infant acts according to the reinforcement learning algorithm described above.

Every 300,000 steps, several experiments are run on the model, replicating the experimental setups used by different researchers. Learning is "frozen" during testing, so that weight values do not change with actions and rewards. This is required to avoid biases towards the test setups, because the experiments are repeated a large number of times.

III.A.5 Parameter setting

This section describes the default parameter values of the model. These are the values used in the next chapters unless otherwise indicated. Table III.1 summarizes the parameters and their settings. These values were made as naturalistic as possible, in some cases using values cited in experiment descriptions.

Environment Modeling parameters: These parameters were set to simulate a naturalistic environment where caregiver and infant interact with each other in a fairly dynamic environment. This is similar to other models from the MESA project (Carlson and Triesch, 2003; Triesch, Teuscher, Deák, and Carlson, 2006b; Lau and Triesch, 2004), and based on assumptions about a structured environment as described in Fasel et al. (2002).

Φ_I, the infant's saliency, is set to a value of 4.0. Φ_C, the caregiver's saliency, is set to a value of 4.0 (with a value of 2.0 when the caregiver is looking sideways). $\bar{\Phi}_o$, the average object saliency, is set to 1.0. This makes the infant and caregiver above-average objects of interest. The caregiver's saliency is given a high value because newborns preferentially orient towards faces (Johnson & Dziurawiec, 1991; Valenza et al., 1996), and because caregivers provide social contingency, which is preferred by infants (Murray and Trevarthen, 1985). With this parameter setting, most of the objects will be less salient than the caregiver or infant, with the possibility of having objects that are more interesting.

The average number of objects, $\bar{N}_o$, is set to 4, for a reasonably rich environment. This value can be set lower (but not to 0) or higher without significant differences in results. The object placement spread, σ_o, is set to 1.0 m. This samples all four depth ranges in the infant's visual system with roughly the same frequency. These parameter values simulate a setting such as a nursery where objects are placed around the infant for it to play with, but with some objects like walls, doors, desks, or chairs far away.

$\bar{T}_{present}$ is set to 60 seconds, $\bar{T}_{absent}$ to 60 seconds, and $\bar{T}_{objects}$ is set to 5 seconds. This models a fairly dynamic environment, with typical object displacements such as the caregiver manipulating a toy in front of the infant while playing or teaching, or the infant itself manipulating the objects. Having the caregiver present half the time simulates the substantial time involved in child rearing when the infant is awake, which includes activities with face-to-face interaction between infant and caregiver, such as feeding and playing, when the infant has access to both the caregiver's face as well as objects being manipulated. Lowering these values can make gaze following be learned more quickly, but if the environment is made to be too dynamic, then the model will have trouble learning to follow gaze. This, because it becomes more likely that as soon as the infant turns away from the caregiver the object moves, making the reference invalid.

Infant visual system parameters: The parameters for the infant visual system were set to simulate a naturalistic setting. Habituation, foveation, and a limited memory make the infant look at the caregiver often for clues about object locations.

FOV, the infant's field of view, is set to 180°, simulating the human visual system. Habituation's τ_H and α_H are set to 2.5 and 1.0 respectively, resulting in almost complete habituation after about 5 seconds. The memory decay factor d is set to 0.5, resulting in all memory traces being cleared after about 5 seconds.

The initial value of σ_H ($\sigma_{H_{initial}}$) is set to 50°. σ_H decrements 5° ($\sigma_{H_{step}}$) every 300,000 time steps, reaching a final value ($\sigma_{H_{final}}$) of 1°. The corresponding values for $\sigma_{E_{initial}}$, $\sigma_{E_{step}}$, and $\sigma_{E_{final}}$ are 50°, 2°, and 1°. This corresponds to an eye direction signal more difficult to interpret than the head direction cue, the eyes being smaller than the head, and allows us to replicate experiments where the value of the other's eye direction is learned slower than that of the head direction. These settings are important to replicate a gradual incorporation of the eyes direction cues. If the final value is reached too quickly (large values of $\sigma_{H_{step}}$ and $\sigma_{E_{step}}$), then the eye direction cue will be incorporated too soon into gaze following.

Reinforcement Learning parameters: In general, these parameters are set so that learning can take place fast, but not so fast that learning becomes unstable.

The learning rate η is set to 0.005 for smooth learning. Making this value higher will cause the learning to become unstable, and is not recommended. The discount factor γ is set to 0.1. The "inverse temperature" parameter β is set to 30, resulting in a high level of exploration early on, and a fairly greedy selection afterwards, as the weight values of $\mathbf{w}$ and $\mathbf{M}$ increase through learning. All elements of $\mathbf{M}$ and $\mathbf{w}$ are initialized to zero, reflecting an absence of previous experience with saliencies and gaze, and of any innate gaze following abilities.

Table III.1: Overview of model parameters, their allowed ranges and default values.

Symbol	Explanation	Range	Default
Environment modeling			
Φ_I	Infant's saliency	$(-\infty, \infty)$	4.0
Φ_C	Caregiver's saliency	$(-\infty, \infty)$	4.0
$\bar{\Phi}_O$	Average object saliency	$(-\infty, \infty)$	1.0
$\bar{N}_o$	Average number of objects	$[0, \infty)$	4
σ_o	Object placement spread around infant	$[0, \infty)$	1.0 m
$\bar{T}_{present}$	Average caregiver interaction interval	$[0, \infty)$	60 s
$\bar{T}_{absent}$	Average caregiver absence interval	$[0, \infty)$	60 s
$\bar{T}_{objects}$	Average object replacement interval	$[0, \infty)$	5 s
Infant visual system			
FOV	Size of field of view	[0°, 360°]	180°
σ_H	Head direction perception fuzzyiness	$(0°, \infty)$	
$\sigma_{H_{initial}}$	Initial σ_H value	$(0°, \infty)$	50°
$\sigma_{H_{final}}$	Final σ_H value	$(0°, \infty)$	1°
$\sigma_{H_{step}}$	Decr.t in σ_H per 300,000 time steps	$[0°, \infty)$	5°
σ_E	Eyes direction perception fuzzyiness	$(0°, \infty)$	
$\sigma_{E_{initial}}$	Initial σ_E value	$(0°, \infty)$	50°
$\sigma_{E_{final}}$	Final σ_E value	$(0°, \infty)$	1°
$\sigma_{E_{step}}$	Decr. in σ_E per 300,000 time steps	$[0°, \infty)$	2°
τ_H	Habituation rate	$[0, \infty)$	2.5
α_H	Target of habituation	$[1.0, \infty)$	1.0
d	Memory decay factor	[0,1]	0.5
Reinf. Learning			
η	Learning rate	$[0, \infty)$	0.005
γ	Discount factor	$[0, \infty)$	0.1
β	Inverse temperature	$[0, \infty)$	30

On using a single set of parameters: The model exhibits two characteristics that make it appealing: First, a single parameter specification is sufficient to replicate a wide variety of gaze following experiment results, as described in the next chapters. This leads to a stronger claim of fitting the data than the alternative method of using different parameter settings for different experiments of the same phenomena (Roberts and Pashler, 2000). Second, the model can still replicate the experiment results even with reasonable modifications to these parameters. For example, the value of the caregiver's saliency (Φ_C) does not need to be exactly 4.0. Any value greater than zero will result in gaze following learning (see related experiment in chapter VI).

It should be noted that many aspects of the model, such as the representation of space, or the different head and eye cues, were introduced because of a desire to replicate as many experiments as possible. And while simpler versions of the model could be used to drive the point for different experiments (for example, the limited field of view is not necessary to replicate experiments investigating the different effect of head and eye cues), there is a value in having a single model with a single set of parameters. In particular, the leap to its implementation as a robotic system should be easier to make.

As more experiments were replicated during the development of the model, its complexity grew. For example, habituation and foveation were not included in the first versions of the model. And having some variables such as the number of objects in the room or the saliency of objects be drawn from a probability distribution (instead of fixing the number to a certain value) proved to be useful to eliminate any biases in learning towards any particular parameter value.

III.A.6 Testing gaze following

To show that the model effectively learns to follow gaze, we used Scaife and Bruner's (Scaife and Bruner, 1975) original experimental setup along with Corkum & Moore's scoring function (1995), which is widely used: Infant and experimenter start in the middle of the room, with the infant looking at the experimenter, and the experimenter turning to the right at 90° from the midline, for 6 time steps (6 seconds). During the trial it is noted whether the model infant a) looks at the *correct* side of the room (i.e. at the side of the room to which the experimenter turned, resulting in a score of +1 for the trial, b) looks at the *incorrect* side of the room, resulting in a score of -1, or c) does not turn gaze, resulting in a score of 0. Scores are averaged over the number of trials that the experimental setup is repeated for, so that a score of 1.0 indicates perfect gaze following.[1]

Fig III.6 shows the result. It shows how the gaze following score increases as the model learns, reaching a high score after about 3,000,000 time steps (scores significantly above zero are considered an indicator of gaze following).

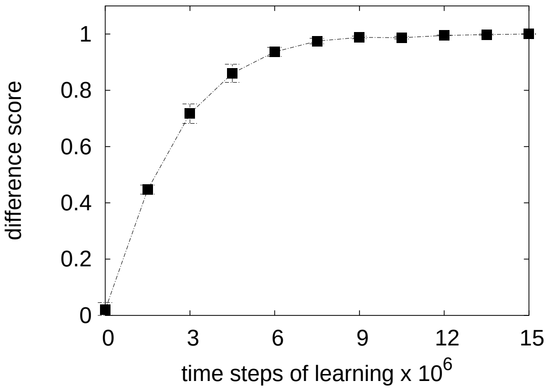


Figure III.6: Measuring basic gaze following performance. Errors bars indicate standard errors after 5 repetitions.

III.B Understanding the Model

This section gives some details about how the model learns to follow gaze.

III.B.1 Learning to look at saliencies

The first thing the model learns is to look preferentially at locations with high saliency. Although simple, this relationship between the saliency of a location and the expected reward resulting from looking at that location is not assumed

[1] We further adapted Corkum and Moore's scoring method: instead of adding the score of four trials to compute the final score, we averaged over however many trials were done. Since Corkum and Moore repeated their experiment four times for each infant, a perfect score for them was 4.

by the model, but instead must be learned. The diagonal at the left part of $\mathbf{M}$ in Fig. III.5 shows this relationship after learning has taken place: a saliency at location i (s_i) will mostly add, through $\mathbf{M}$, to the element of m corresponding to action i (m_i), increasing the probability of looking at location i. Before learning, all values are zero. Human infants also take time to develop this behavior, consistently saccading to stimulus contours at 14 weeks but not at 2 weeks of age (Bronson, 1990).

III.B.2 Learning to follow gaze

Later on, the model learns to follow gaze: The diagonals on the right part of $\mathbf{M}$ in Fig. III.5 cause activations in h_i or e_i, which result from the caregiver turning/looking in a particular direction, to add to elements of m along that direction.

This phase takes longer to learn because there is a one-to-many relationship between a caregiver looking direction (elements in $\mathbf{h}$ and $\mathbf{e}$) and the actions (elements in $\mathbf{m}$) corresponding to looking at locations along the caregiver's corresponding line of sight (see Fig. III.7). Additionally, the model loses opportunities to learn to follow gaze in the times when the caregiver is not present.

III.B.3 Bottom-up and top-down visual search

In visual search, bottom-up processing refers to the preference for looking at locations with salient objects, such as a red balloon among black ones, or a light source in a scene. Although during infancy our visual search is mostly driven by bottom-up processing, with time we learn to look at positions that might not be very salient, either hoping to find something of interest, or because we are performing a task not related to visual salience, such as looking for a person in a crowd, counting objects on a table, or planning a reaching movement.

In the model, bottom-up visual search corresponds to looking at locations where salient objects are. Looking at locations that the caregiver might be looking

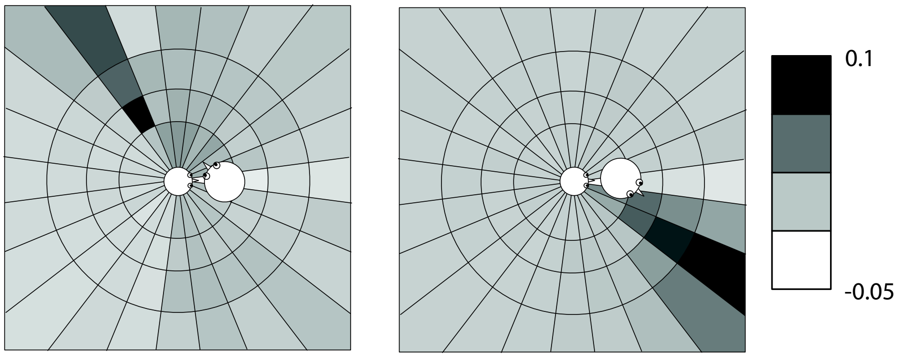


Figure III.7: Illustration of connection weights from inputs **h** and **e** to vector **m** after gaze following is learned. Shown are two different caregiver head/eye directions and the corresponding activations in **m** (since the caregiver's head and eyes are aligned in these examples, the values of **h** and **e** are the same).

at is the top-down part, because it involves decoding the caregiver's head and eye direction in order to find out the possible locations to look at.

To see how top-down visual search is gradually integrated in the model, an experimental setup was created, as depicted in Fig. III.8: Trials start with the infant looking at the experimenter, and the experimenter looking to the right at 60° from the midline, towards an object (object A). Another object (object B) is positioned on the opposite of the room from object A. Object A's saliency is 80% of object B's. Trials last 6 seconds, after which it is noted what object the infant turns gaze to. If bottom-up influences are stronger than top-down influences, the infant will tend to look at object B, which is more salient but not being looked at. But as top-down influences are incorporated, the likelihood that the infant will disregard object B's saliency in favor of following the experimenter's gaze to object A will increase.

Trials were repeated 200 times, 100 for the setup shown in Fig. III.8, and 100 for a "mirror setup", where objects A and B are swapped but with the caregiver still looking at object A. Fig. III.9 shows the percentage trials in which the infant either looks at object A, object B, or at other (empty) locations. As

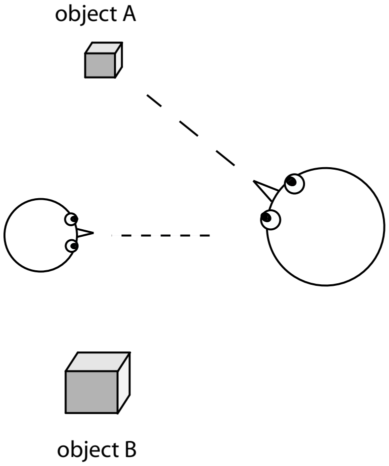


Figure III.8: Bottom-up and top-down visual attention integration: experimental setup. Object A is only 80% as salient as object B.

the value of bottom-up cues is learned, the infant preferably looks at object B, which is more salient. But as the infant learns to follow gaze, it starts to look more at object A, which is less salient but being looked at by the experimenter. (The results for 0 time steps correspond to random action selection, where the model infant is mainly exploring.) This shows a gradual integration of top-down attention into earlier bottom-up attention.

III.C Discussion

The model presented provides a new computational account of gaze following, based on the MESA project's Basic Set Hypothesis (Fasel et al., 2002). Its usefulness as a model of the mechanism behind gaze following will become more evident as it is used in the next chapters to replicate various experiments carried out in gaze following research.

III.D Acknowledgements

The text of this chapter, in part, is a reprint of the material in:

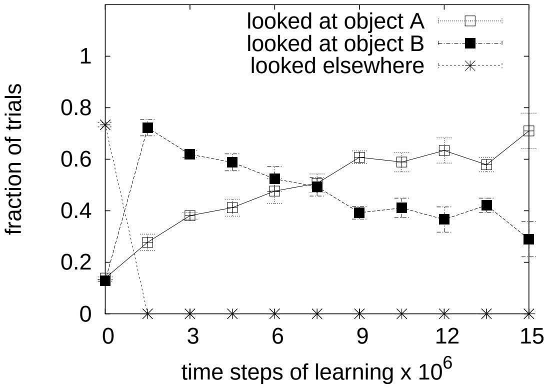


Figure III.9: Bottom-up and top-down visual attention integration: results. Percentage trials where infant looks at object A, object B., or at other (empty) locations. From setup depicted in Fig. III.8. Error bars indicate standard error after 5 repetitions.

H. Jasso, J. Triesch, C. Teuscher, and G. O. Deák. A reinforcement learning model explains the development of gaze following. In *Proceedings of the Seventh International Conference on Cognitive Modeling (ICCM 2006)*, Trieste, Italy, April 2006.

The dissertation author was the primary investigator and author of this paper.

IV

Refining Geometric and Representational Aspects of Gaze Following

George Butterworth, Edward Cochran, and Nicholas Jarrett wanted to show how gaze following could happen without a sophisticated understanding of the other person's attention. They believed that attentional mechanisms played a large part instead. They presented a nativistic interpretation of gaze following, consisting of three such mechanisms that develop during the first 18 months of age. These were used to explain why, although younger infants will only follow gaze to objects within their field of view (in front of them, when facing the experimenter), older infants also follow gaze to objects outside their field of view (behind them, when facing the experimenter). Also, why younger infants but not older ones turn to the correct side of the room as indicated by the experimenter's head or eyes direction, but often end up looking at other objects not looked at by the experimenter but on the correct side of the room.

The first section of this chapter describes the different experiments carried out to measure these spatial refinements, and their interpretation of the results as revealing three underlying mechanisms behind gaze following (Butterworth and Cochran, 1980; Butterworth and Jarrett, 1991; Butterworth, 1995). The second section shows how the model presented in the previous chapter also displays these transitions, but using a single mechanism.

IV.A Butterworth's Three-Mechanism Interpretation

This section describes the experiments done by Butterworth and colleagues to measure spatial refinements of gaze following.

IV.A.1 Two-target setting and results

Butterworth and colleagues used Scaife and Bruner's basic setup (1975), with some modifications: In one set of experiments, the *two-target setting*, (Butterworth and Cochran, 1980; Butterworth and Jarrett, 1991) targets were positioned two at a time, one on each side of the room along the wall, as shown in Fig. IV.1 a). In the first three variations of the experiment (top row), the targets were set at either 30°, 60°, or 90° from the infant's midline. These correspond to targets *within the infant's field of view* (when looking at the experimenter). In the other two variations (bottom row), the targets were set at either 120° or 150°. These correspond to targets *outside the infant's field of view.*

A *correct* response was defined as looking at plus or minus 30° around the correct target. A *wrong* response was defined as looking elsewhere but on the same side of the room as the correct target. *Non-codable* responses were those where the infant made no response or when a response was made in other than the horizontal plane (e.g. looking up). A final category of responses included cases where the infant looked at the opposite side to the mother's direction of gaze (e.g. mother looking to her right and infant looking to his/her right). These were infrequent and omitted in any score calculations. Accuracy was measured as the number of correct responses over the sum of correct plus wrong responses.

Each trial type was repeated twice, once for each side of the room. The results are presented in Fig. IV.2 left, in Fig. IV.3 a), and in Table 1 in the

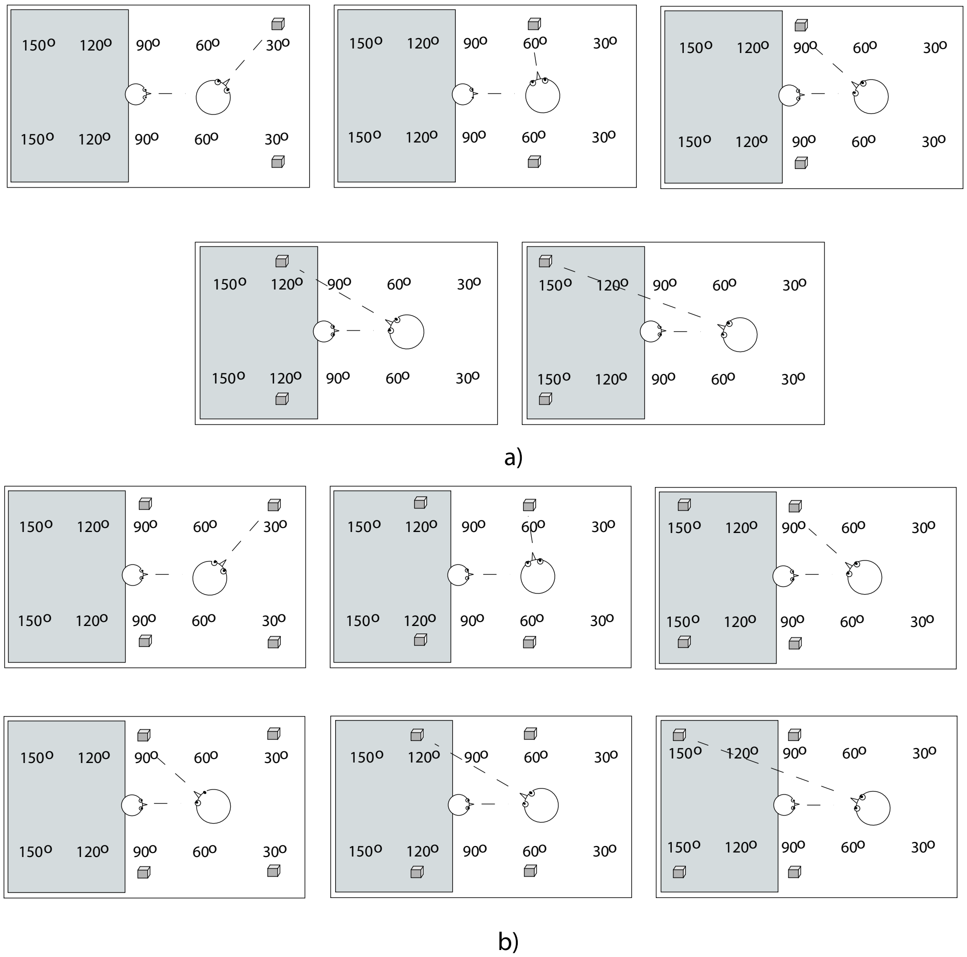


Figure IV.1: a) Butterworth's two-target setting. Gray area represents space outside the infant's field of view. Top row: Target *within the infant's field of view*, at 30° (left), 60° (middle), or 90° (right) from the infant's midline. Bottom row: Target *outside the infant's field of view*, at 120° (left), or 150° (right). b) Four-target setting. Top row: Target *first along the scan path*, at 30° (left), 60° (middle), or 90° (right). Bottom row: Target *second along the scan path*, at either 90° (left), 120° (middle), or 150° (right).

Appendix. [1]

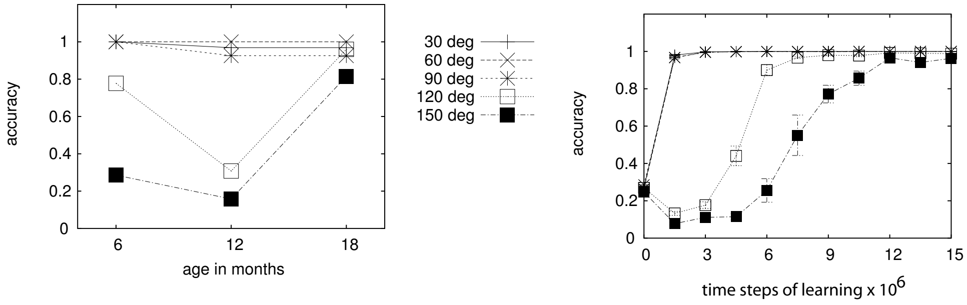


Figure IV.2: Results for the two-target setting. Left: Butterworth's results. Right: Model's results (error bars indicate standard errors after 5 repetitions).

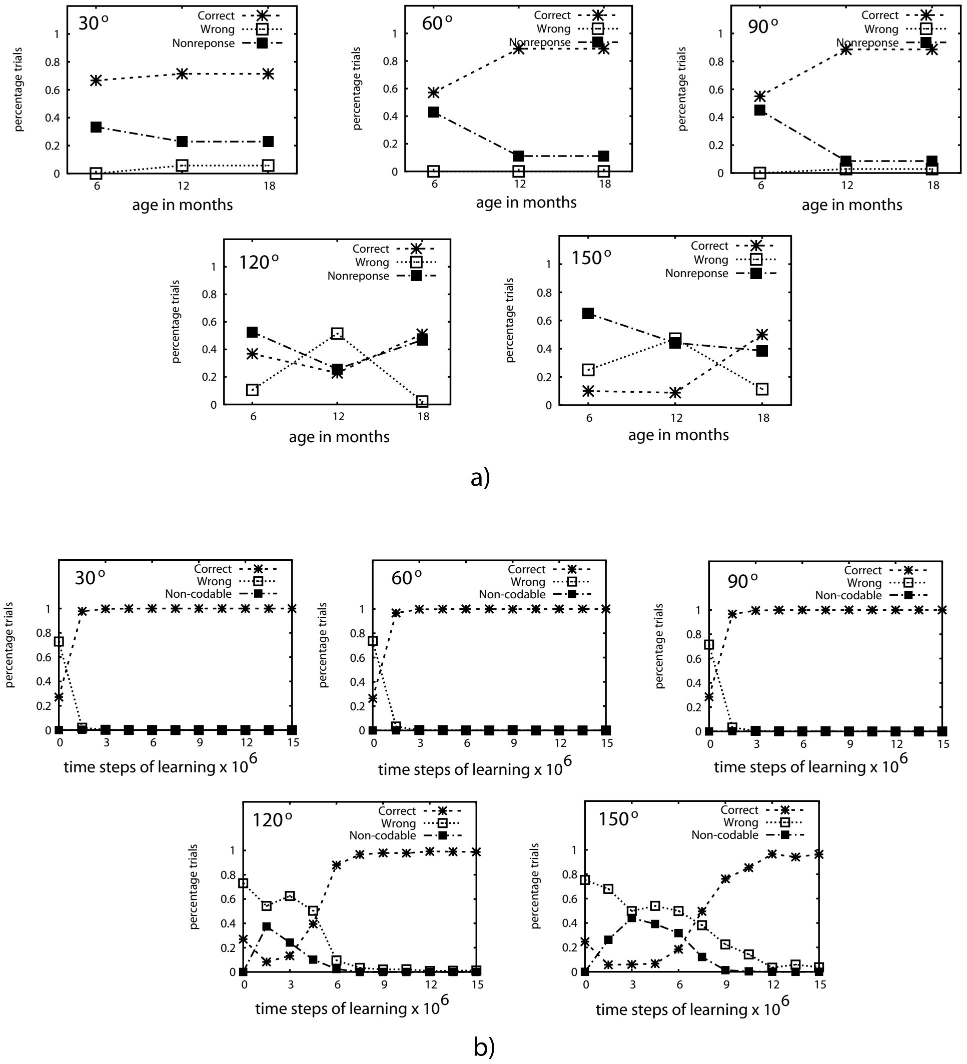
These results show that infants age 6 months and older turn when the target is within their field of view (angles of 30°, 60°, and 90°). But not until 18 months of age will they consistently follow gaze to targets outside their field of view (targets at 120° and 150° from their midline).

IV.A.2 Four-target setting and results

In another experimental setup (Butterworth and Cochran, 1980; Butterworth and Jarrett, 1991), the *four-target setting*, targets were positioned four at a time, two on each side of the room along the wall, separated 60° from each other, as shown in Fig. IV.1 b). In the first three variations of the experiment (top row), the correct target is *first along the scan path* when compared to the other target in the same side of the room, while in the last three (bottom row) the target is *second along the scan path.*

In this setting, *wrong* responses were measured as looking at plus or minus 30° around the incorrect target on the same side of the room. The other responses

[1] Results for 6 months of age correspond to experiment 3c in (Butterworth and Jarrett, 1991) (12 subjects). Results for 12 months of age correspond experiment 2 in Butterworth and Cochran (1980) (18 subjects). Results for 18 months of age, 150° and 180° correspond to experiment 2 in Butterworth and Jarrett (1991) (18 subjects). No experiments were carried out for 30°, 60°, and 90° at 18 months of age, because the infants had already near-perfect results for these target positions since 6 months of age. But for display purposes for Fig. IV.2 and Fig. IV.3, results for 12 months of age for these target positions were repeated for 18 months of age.



a)

b)

Figure IV.3: Details (correct, wrong, and non-codable responses) for the two-target setting for individual target positions. a) Butterworth's results. b) Model's results.

(*correct*, *non-codable*, and the omitted ones), as well as the accuracy measure, were the same as in the two-target setting.

Each trial type was repeated twice, once for each side of the room. The results are presented in Fig. IV.4 left, in Fig. IV.5 a), and in Table 2 in the Appendix. [2]

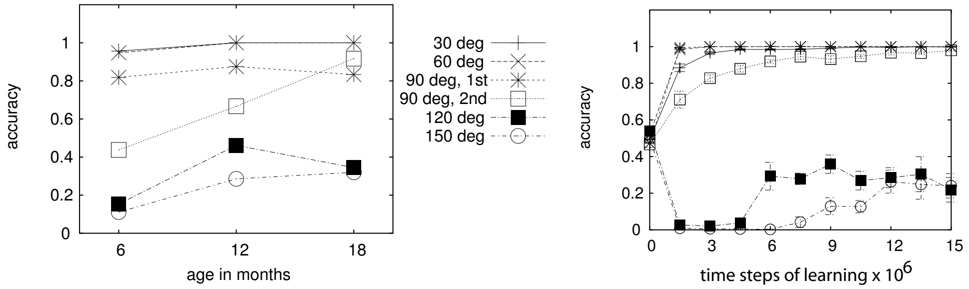


Figure IV.4: Results for the four-target setting. Left: Butterworth's results. Right: Model's results (error bars indicate standard errors after 5 repetitions).

Results show that at all ages infants reliably follow gaze when the correct target is positioned first along the scan path. When the correct target is second along the scan path and within the infant's field of view (Fig. IV.1 b), bottom left), 6-month-olds stop at the distracter object about half the time, 12-month-olds disregard the distracter more often, and 18-month-olds follow gaze correctly, disregarding the distracter. At no age did infants reliably follow gaze when the correct target was second along the scan path and outside the field of view (Fig. IV.1 b), bottom middle and bottom right).

IV.A.3 Ecological, geometric, and representational stages

According to Butterworth and colleagues (Butterworth and Cochran, 1980; Butterworth and Jarrett, 1991; Butterworth and Grover, 1988, 1990; Butterworth, 1991, 1995), this gradual improvement in gaze following expertise can

[2] Results correspond to experiment 1 in Butterworth and Jarrett (1991) (18 subjects). A subset of these experiments was tested in Butterworth and Cochran (1980) and repeated in Butterworth and Jarrett (1991); since results were similar, only the most recent ones (i.e. Butterworth and Jarrett (1991)) are presented.

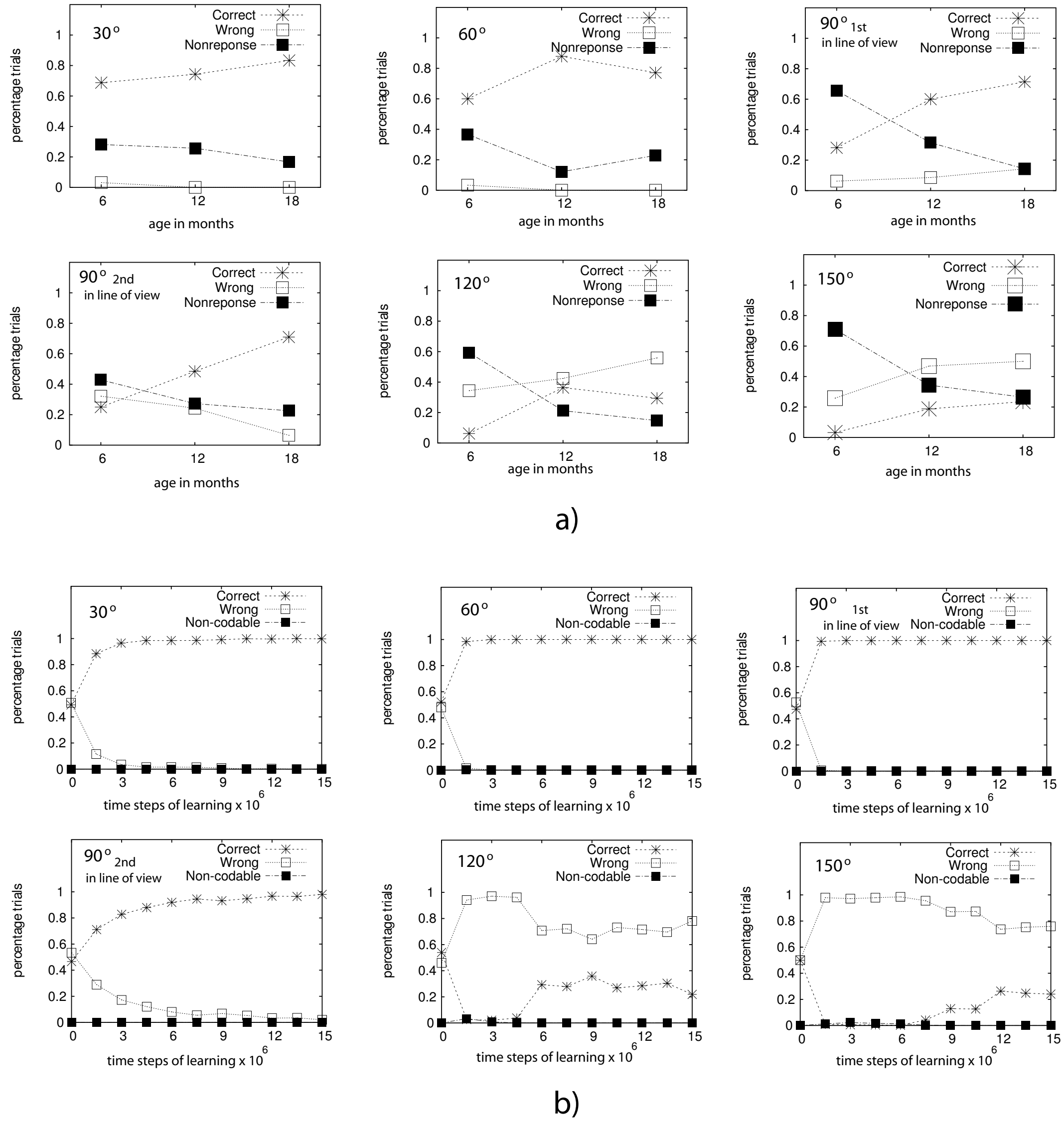


Figure IV.5: Details (correct, wrong, and non-codable responses) for the four-target setting for individual target positions. a) Butterworth's results. b) Model's results.

be explained as the product of three sensorimotor/cognitive mechanisms appearing successively during the infant's development, improving its mastery of spatial knowledge. They postulated the mechanisms as follows: The first appears at around six months of age. At this age infants look to the correct side of the room, but if there is more than one (identical) target on that side of the room, the infant cannot tell on the basis of the experimenter's action alone which of the two the experimenter is attending to. Although they are accurate in locating the target when it is first along their path of scanning from the experimenter, they are at chance level when the correct target is second along the scan path. Also, if the experimenter looks at targets behind the infant (outside its field of view), the infant either fixates a target in front and within the visual field or does not respond. Butterworth attributes this to a basic inability to link the experimenter's signal to the space outside its immediate visual field. This was not caused by an inability of infants to turn behind them, because they would do so on first being seated in the laboratory or in response to some inadvertent noise. Neither was it the small change in the experimenter's head direction that caused infants not to look back (Butterworth and Jarrett, 1991). Additionally, Grover (1988) showed that if the correct target is very salient then infants always turned to look at it. For these reasons, Butterworth believed that the earliest mechanism of joint visual attention was an *ecological* one, where the differentiated structure of the natural environment is what completes for the infant the communicative function of the adult's signal.

By 12 months of age infants begin to localize the target correctly when it is first or second along the scan path. Butterworth calls this new ability the *geometric* mechanism. This mechanism involves extrapolation of an invisible line between the experimenter and the referent of her gaze. The experimenter's change of gaze therefore signals both the direction and the location in which to look. But at this age infants still will not search for targets located behind them. Instead, they turn to scan to about 40 degrees of visual angle and give up the search when they fail to encounter a target. This made Butterworth believe that the geometric mechanism is restricted to the infant's perceived space.

By 18 months of age, infants are as accurate when the correct target is first along their scan path from the experimenter, as when it is the second target they encounter, suggesting that the geometric mechanism is fully available. Also, at this age infants will search behind them, although only when their visual field is empty of targets. This led Butteworth to postulate the development of a third, *representational*, spatial mechanism for controlling joint visual attention, a mechanism based on an understanding of being contained in space.

IV.B A Single-Mechanism Interpretation

This section describes how the model also transitions through the stages described above, thus unifying them into a single learning account.

IV.B.1 Experiment simulation

Butterworth's experiments were simulated, using an object saliency (Φ_o) of 1, corresponding to objects of an average saliency. Each trial was run for 6 time steps, equivalent to the 6 seconds used in the experimental setups described above.

For the two-target setting, the accuracy was calculated as follows: *Correct* responses are those where the model's attention (heading) shifts from initially looking at the experimenter (a heading of 1) directly to looking at the target, or to a heading immediately to the right or the left of the target (heading of 2, 3, or 4 for a target at 30° to the left, heading of 22, 23, or 24 for a target at 30° to the right, heading of 4, 5, or 6 for a target at 60° to the left, and so on). This corresponds to looking at the target plus or minus 22.5°, which is slightly more strict than the plus or minus 30° used by Butterworth. *Wrong* responses are those where the infant looks at the correct side of the room but misses the target. If during the trial the infant does not shift gaze, the response is considered *Non-codable*. Responses where the infant looks at the wrong side of the room are omitted, as Butterworth did (these responses were also infrequent in the model).

For the four-target setting, the accuracy was calculated as follows: *Correct* and *Non-codable* responses are defined as above. *Wrong* responses are defined as for the correct responses, but with respect to the distracter on the same side of the room as the correct target. Responses where the infant looks at the wrong side of the room are also omitted. Added to this category are responses where the infant looks at the correct side of the room but at no targets. This type of response is not explicitly considered by Butterworth, supposedly because infants rarely turn gaze towards empty space. For example, as Fig. IV.5 a) shows, in the cases where there is a target within the infant's field of view (30°, 60°, and 90°) the number of wrong responses (i.e. looking at empty space) is close to zero.

In real experimental setups, only the infant's direction of attention is scored, and not the depth of attention, which is difficult if not impossible to obtain. For the same reason, scoring in the model uses only the infant's direction (φ_I).

Each 1,500,000 time steps all the experimental variations were repeated 200 times, 100 times for each side of the room. The complete process was repeated 5 times. Fig. IV.2 right and Fig. IV.3 b) show the results for the two-target, and Fig. IV.4 right and Fig. IV.5 b) show the results for the four-target setting. At 0 time steps, no learning has taken place so that the results reflect random actions. These random actions cause almost no non-codable responses because during the six seconds of the trials the model is likely to turn away from the experimenter as it chooses a random action on each step. In the 2-target settings, these random actions result in mostly Wrong responses and a small number of Correct responses, because within the correct side of the room most headings correspond to empty space.

IV.B.2 An alternative view of the representational stage

At 1,500,000 time steps, the accuracy for the 30°, 60°, and 90° variations in the two-target setting (corresponding to targets within the infant's field of view) is already close to 100%, while accuracy for the 120° and 150° variations (corresponding to targets outside the infant's field of view) is close to 0% (see Fig. IV.3 b)). This corresponds to 6-month-olds in Butterworth's experiments, where gaze is correctly followed to targets within the field of view, but not if they are outside the field of view.

By 9,000,000 time steps, the accuracy of gaze following to targets outside the infant's field of view is close to 100% for the 120° setting, and about 80% for the 150° setting. These values correspond to the accuracies of 18 month olds in Butterworth's experiments, where, according to Butterworth, a *representational* mechanism has been incorporated.

This transition from following gaze only to targets within the field of view to also include targets outside the field of view can be explained as follows: Since the model learns to look at salient locations before it learns to follow gaze (see previous chapter for a discussion of this), this biases the initial learning of gaze following to objects within the field of view: Initially, the high saliency of the caregiver (4 times that of the targets) will attract the infant model's attention in that direction (heading of 1, or $\varphi_I = 0°$) often. After habituating to the experimenter, the infant model is likely to look at objects within the field of view because of their visual saliency. This happens independently of the experimenter's head/eyes direction. But when the object the infant looks at corresponds to the one that the experimenter is looking at, this causes the model to learn an association between the experimenter's head/eye direction and the value of looking at objects in that location. Although exploration will help the model learn this association, it is not as important as in the case of learning to follow gaze to objects outside the field of view. There, the model must ignore the objects within the field of view and turn back.

IV.B.3 An alternative view of the geometric stage

At 1,500,000 time steps, the accuracy of the 30°, 60°, and 90°-*first-in-line-of-view* in the four-target setting (corresponding all to the target being first in line of view) is already close to 100% (see Fig. IV.5 b)). For the case of 90°-*second-in-line-of-view*, the infant model looks at the two targets with similar probability (in (Butterworth and Cochran, 1980), this value was 30%, while in (Butterworth and Jarrett, 1991), it was 50%). When the target is second in line of view *and* outside the field of view (120° and 150° cases), the accuracy drops to close to zero. Therefore, this corresponds to infants of 6 months of age.

From 3,000,000 to 6,0000 training steps, the accuracy of the 90°-*second-in-line-of-view* setting progressively improves, while the accuracy for the other settings remains the same. This corresponds to infants of 12 months of age, where, according to Butterworth, a *geometric* mechanism is being incorporated into gaze following.

At 9,000,000 training steps, the accuracy of 90°-*second-in-line-of-view* is close to 100%. This corresponds to infants of 18 months of age, where the geometric mechanism is fully incorporated.

The model's progression in the 30°/90°-*second-in-line-of-view* setup can be explained as follows: At first, foveation causes the 30° target to be more salient than the one at 90°, when their intrinsic saliency is the same. But since the experimenter is looking at the 90° target, this balances the selection between the two targets. If targets of higher saliency are used, the model selects the 30° target over the 90° target, and if less salient targets are used, the 90° target is preferred. Therefore, the target that is selected is defined by the relative influences of $\mathbf{s}$ and $\mathbf{h}/\mathbf{e}$.

What causes the transition to selecting the 90° over the 30° target is the decay in σ_H and σ_E, the parameters that define how "fuzzy" the signal of the experimenter's head/eyes direction is reflected in $\mathbf{h}/\mathbf{e}$ (see description of infant visual system in the previous chapter). As σ_H and σ_E diminish in value, the values of **h** and **e** reflect the experimenter's head and eye direction more sharply. This also results, through learning, in narrower and "higher" diagonals in the sections of **M** corresponding to **h** and **e** (see Fig. III.5 and Fig. III.7). Overall, this leads to a stronger effect with time of the head and eyes direction cues, which eventually win over the saliency of the 30° target, resulting in the infant looking at the 90° target.

It should be noted that the accuracy at 12 months in Butterworth and Cochran (1980) is around 30%, contrasting with the results in Butterworth and Jarrett (1991) of around 70% at 12 months, and 50% at 6 months for a similar setup. The model explains the difference in results for the same experiment as the product of a 'delicate balance' between the foveation and the target saliency, making it sensitive to target saliency.

The model also explains why in the 120° and 150° settings gaze is not followed: the saliency of the distracter is enough to override the effect of the experimenter's head and eyes direction, because of the foveation offset, which does not let saliency drop to values too close to zero when the distracter is positioned at 30° and 60°, respectively. This is illustrated by repeating the 4-target setting with target at 120° (Fig. IV.1 b), bottom row, middle) with different values for the object saliencies. Although both target and distracters vary their saliency, the target (and the distracter that mirrors the target at the other side of the room) is not visible to the infant. The saliency of the visible distractors is what changes the results of the experiments, as shown in Fig. IV.6: decreasing object saliencies helps the infant disregard the distracter (object saliencies of 0 and 0.5), while increasing the object saliencies makes the infant look at the distracter instead (object saliencies of 1.5 and 2.0). This effect was observed by Grover (1988), who noted an almost 100 per cent likelihood of attending to the first target along the scan path (the distracter on the same side of the room as the target) when the saliency of targets was increased by setting them in motion.

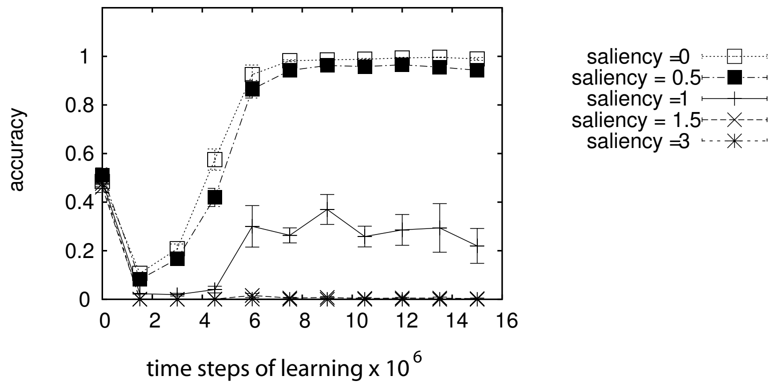


Figure IV.6: Effect of varying object saliencies in the 4-target, 120° setting: Increasing the saliency causes the infant to look at the distracter, resulting in a low score (accuracy) for this setting. But with a low target saliency, the infant is able to disregard the distracter and achieve a high accuracy. Note that the result for object saliency = 1 is the same as that of Fig. IV.4 right, 120°.

IV.C Discussion

Originally, these experiments were carried out by Butterworth to show that gaze following in infants does not necessarily indicate that they entertain a 'theory' that other people have minds (this notion is explored further in the next chapter). The experiments therefore highlight spatial aspects of gaze following, to indicate how attentional mechanisms, combined with a social context (i.e. interactions with the experimenter), could explain gaze following. The model presented is in the same spirit as Butterworth's explanation, with attentional mechanisms developing within a social context. The main difference is that our model gives a central role to learning: It explains the gradual improvement of gaze following skills (from the ecological, to the geometric, to the representational stages), using a single mechanism. Butterworth's three-mechanism explanation is less parsimonious in that it requires additional explanations of how the mechanisms are genetically encoded and how they are integrated during development.

IV.D Acknowledgements

The text of this chapter, in part, is a reprint of the material in:

H. Jasso, J. Triesch, C. Teuscher, and G. O. Deák. A reinforcement learning model explains the development of gaze following. In *Proceedings of the Seventh International Conference on Cognitive Modeling (ICCM 2006)*, Trieste, Italy, April 2006.

The dissertation author was the primary investigator and author of this paper.

V

Gaze Following and the Understanding of Other's Attention

A recurring question in gaze following studies is whether infants follow gaze because they understand about the other person's visual attention (the 'mentalist' position), or simply because they have learned the value of head and eye direction cues as indicators of locations of interesting objects (the 'attentional' position). If the former is true, then they can be said to possess at least the beginnings of a 'theory-of-mind', the capacity to impute mental states such as beliefs and intentions to others (Premack and Woodruff, 1978).

The first section of this chapter describes this debate in more detail. The second section reviews the experiments related to this debate. These experiments have focused on the transition from using head direction cues to using eye direction cues in gaze following, where the latter is believed to be a true indication of theory-of-mind abilities by some authors. The third section shows how the model can be used to replicate this transition even when both cues are available from the start. This unifies mentalist and attentional positions into a learning account.

V.A Mentalist vs. Attentional Interpretations of Gaze Following

There are two positions with respect to the relationship between gaze following and theory-of-mind. The first position is the 'mentalist' view of gaze following (Caron, Butler, and Brooks, 2002). Baron-Cohen (1991) and (Bretherton, 1991), for example, argue that gaze following in infants happens as they realize that the other person is looking at something, and try to adjust their gaze to match it with the other's attention. Similarly, Bruner (1995) views innate early social perception and attention skills as necessary to start the learning process leading to joint attention, with a construal of people as agents (i.e. an understanding that human actions are dedicated to attaining ends). (Refer to chapter II for a more detailed explanation of Bruner's position.) Another proponent of this position is Tomasello (1995). He argues that underlying infants' early skills is their emerging understanding of other persons as intentional agents. That is, that other persons attend selectively to certain objects while ignoring others, and that they might intend one to do the same through certain behaviors. (This in turn is based on Gibson & Rader's (1979) notion of attention as intentional perception.) According to Tomasello, gaze following prior to 12 months of age can be considered the product of conditioning. But from 12 months of age on, a qualitatively different gaze following happens, with a real understanding of others as having attention. And while not completely ruling out a conditioning account of gaze following at this age, he considers it unlikely. (Refer to chapter II for a more detailed explanation of Tomasello's position.)

The 'attentional' position, in turn, argues that it is not necessary to attribute such cognitive complexity to infants just because they follow another's gaze (Corkum and Moore, 1995; Perner, 1991; Dunham and Dunham, 1995; Repacholi and Gopnik, 1997; Wellman, 1991). Instead, it might simply result from using another's cues such as head and eyes direction to locate interesting objects in the room, without any need to attribute mental states to the other person. The attentional view of gaze following is in part inspired by observations that infants can, somewhat paradoxically, follow gaze even without knowing some basic facts about the visual perspectives of others, such as that what they themselves see may differ from what others see (Level I perspective taking skill (Flavell, 1974; Moll and Tomasello, 2006)) or that others might see things from a different visual perspective from them (Level II perspective taking skill (Flavell, 1974; Moll and Tomasello, 2006)).

V.B Eye Sensitivity Experiments

This section describes experiments used to argue for either a mentalist or an attentional position of gaze following.

V.B.1 Head vs. eyes experiments

Corkum and Moore (1995) used Scaife & Bruner's original experimental setup (1975), but with some changes in the experimenter's head and eye directions (refer to Fig.V.1):

- In the 'H + E' condition, the experimenter turns both head and eyes towards one side of the room (this is the standard gaze shift used by Scaife & Bruner).
- In the 'H' condition ,only the experimenter's head turns, while the eyes are kept directed towards the infant.
- In the 'E' condition, only the experimenter's eyes turn, while the head is kept directed towards the infant.
- In the 'H - E' condition, the experimenter's head turns to one side of the room, while the eyes turn to the other side (resulting in incongruent cues, a somewhat unnatural setting).

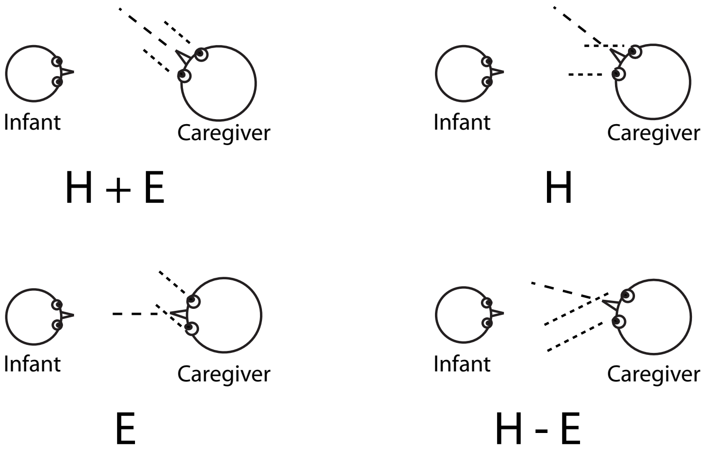


Figure V.1: Experimental conditions set up by Corkum and Moore (1995): 'H + E': Both head and eyes turn. 'H' Only head turns. 'E': Only eyes turn. 'H - E' Head turns in one direction, eyes in the opposite direction.

Each condition was repeated four times, two to the left side and two to the right side. A gaze following score, termed *difference score*, was calculated for each infant, for each condition, by adding the result score of each trial. The result score of each trial is defined as follows: trials where the infant turns to the *correct* side of the room scored a 1, trials where the infant turns to the *incorrect* side of the room scored a -1, and trials where the infant does not turn and instead keeps looking at the caregiver, termed *non-responses*, scored a 0. The "correct side" is defined as the side towards which the experimenter's head or eyes turned. In the 'H - E' condition, this is the side corresponding to the direction of the *head's* turn. Scores in the 'H' and 'H - E' conditions, therefore, measure preference in use of the head direction cue with respect to the eye direction cue. The 'E' condition score, in turn, measures preference of eye direction cues, while in the 'H + E' condition, scores simply measure gaze following.

12 infants for each age group were tested, of ages 6 to 19 months. Results are shown in table V.1 (note that a difference score of 4 represents perfect gaze following).

Table V.1: Results for (Corkum and Moore, 1995).

	Trial Type			
Age (months)	H	E	H + E	H - E
	Matches			
6-7	.917 (.793)	.917 (.669)	.833 (.718)	1.000 (.953)
9-10	1.167 (.835)	1.167 (.718)	1.500 (.905)	1.333 (.492)
12-13	.833 (.835)	.583 (.900)	1.250 (.866)	1.167 (1.267)
15-16	.917 (.900)	.833 (.835)	1.667 (1.231)	.750 (.754)
18-19	.417 (.515)	.583 (.900)	2.000 (1.279)	.500 (.674)
	Mismatches			
6-7	.917 (.793)	1.333 (.888)	1.083 (.996)	.750 (.965)
9-10	1.000 (.853)	.917 (.793)	.917 (.793)	1.000 (.603)
12-13	.500 (.674)	.417 (.515)	.417 (.669)	.583 (.996)
15-16	.417 (.515)	.750 (.965)	.500 (.674)	.750 (.866)
18-19	.750 (.866)	.333 (.492)	.083 (.289)	.417 (.669)

Corkum & Moore's observations were as follows:

- From 6 to 10 months of age, no gaze following was found (no score ('H + E', 'H', 'H - E', 'E') was significantly greater than zero).

- At 12 to 13 months of age, infants seem to be following gaze, but based primarily on head direction (the pooled score based on head direction cues ('H + E', 'H', and 'H - E') was found to be significantly above zero, and no difference in scores was found between any condition ('H + E', 'H', 'H - E', 'E')).

- At 15 months of age, infants now follow gaze ('H + E' score significantly greater than zero), based primarily on head direction ('H + E' score greater than 'E' score; 'H + E' score not significantly higher than 'H' score), but with some sensitivity to eye direction ('H + E' score higher than 'H - E' score).

- At 18 to 19 months of age, evidence for an effect of the eye cue was found, but only with congruent head and eye orientation ('H + E' score higher than the ‘H’, ‘E’, and ‘H - E’ scores; no difference between the ‘H’, ‘E’, and ‘H - E’ scores; and only the ‘H + E’ score was greater than zero).

Corkum and Moore argue that the discrepancy between their results at 18 months of age and that of Lempers (1979) and Butterworth & Jarrett (1991), who found gaze following based on eye direction alone, can be explained by procedural differences: They presented their trials in separate blocks, and this might have enhanced the saliency of the ’E’ trials. In contrast, Corkum & Moore presented trials in such a way that a trial with a strong head signal might come shortly before one with a strong eye signal, diminishing the effect of the eye cue. Also, while Corkum and Moore did not find evidence for a difference between the ‘H + E’ and the ‘H’ variations before 18 months of age, (Caron, Butler, and Brooks, 2002) found a difference at 14 months of age by testing more infants (32 infants, instead of 12).

V.B.2 Eyes open vs. eyes closed experiments

Brooks & Meltzoff (Brooks and Meltzoff, 2002, 2005; Meltzoff and Brooks, 2006) tried an experimental procedure that tested for eye direction while controlling for the head direction cue: Corkum & Moore’s ‘H + E’ variation (referred to as the ‘Eyes Open’ condition now) was compared against a similar setup where the experimenter turned with eyes closed (referred to as the ‘Eyes Closed’ condition).

32 infants per age group were tested. Resulting difference scores are shown in Fig. V.2 left (note that a difference score of 4 represents perfect gaze following) [1]. Starting at 10 months of age, difference scores for the ‘Eyes Open’ condition were significantly higher than for the ’Eyes Closed’ condition.

These results were interpreted as showing a transition from attentional to mentalist gaze following between 9 and 10 months of age.

[1] Results for ages 9, 10, and 11 months are from (Brooks and Meltzoff, 2005); results for ages 12, 14, and 18 months are from (Brooks and Meltzoff, 2002)

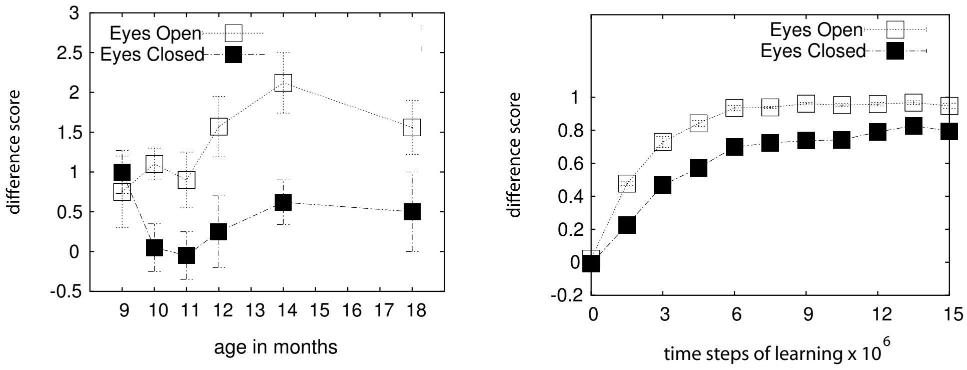


Figure V.2: Left: Results for Brooks & Meltzoff's conditions. Right: Results for the simulation.

V.C Experiment Simulation

In this section, we use the model to simulate the experiments described above.

V.C.1 Simulating head vs. eyes experiments

Fig. V.3 shows the results of using the model to simulate Corkum & Moore's experiments, described above (note that here a difference score of 1 represents perfect gaze following). Details of the results for the different conditions are shown in Fig. V.4.

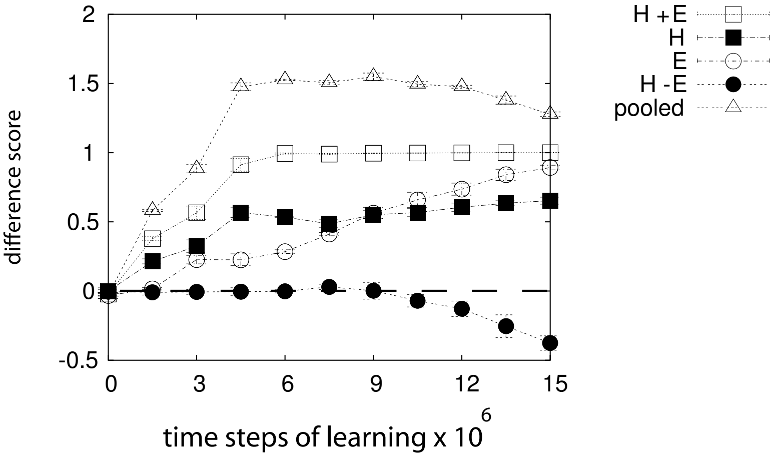


Figure V.3: Results for simulation of conditions depicted in Fig.V.1.

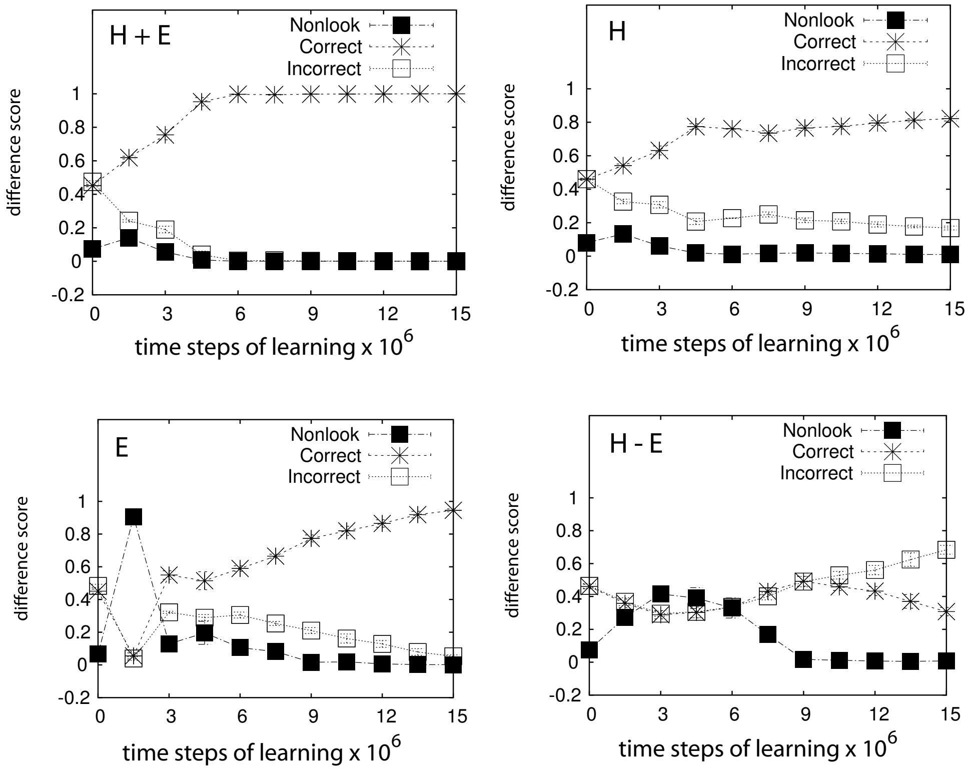


Figure V.4: Result details (non-responses, correct responses, incorrect responses) corresponding to Fig. V.3.

Normal gaze following (‘H + E’ condition) is the first to appear because the infant receives signals to turn from both the head and eyes (**h** as well as **e** add to the values of **m** corresponding to the direction of the experimenter’s head or eyes direction). Gaze following based on eye cues alone (‘E’ condition) is slower to appear because only the eye direction signal is given. And while all ‘E’ scores have non-zero values shortly after learning begins (refer to Fig.V.3), these would be difficult to detect without enough experimental repetitions. Corkum & Moore (1995), therefore, might have not found some of the scores in their experiments to be significantly different from zero or different from each other simply because of insufficient statistical power. Caron, Butler, & Brooks found a difference between the ’H + E’ and the ’H’ condition at 14 months instead of the 18 months measured by Corkum & Moore by increasing the number of experimental repetitions. The simulation supports the idea that the additional repetitions resulted in an earlier measurement of the effect of eye direction cues. It also predicts that with enough repetitions, this effect will be measured at an earlier age. Also, the model predicts that in later ages the eye direction cues will have a stronger effect than head direction cues (difference score for ‘E’ condition higher than the difference score for the ‘H’ condition at the end of the simulation), reversing the earlier trend where head direction cues are stronger (‘H’ higher than ’E’). This is caused by a smaller σ_H than σ_E during early learning, followed by equal values about halfway through learning, and the (small) offset in the head direction with respect the “true” eye direction cue throughout all learning (see chapter III for a description of the behavior of these parameters).

V.C.2 Simulating eyes open vs. eyes closed experiments

Fig.V.2 right and Fig.V.5 show the results of simulating Brooks & Meltzoff’s experiments, described above. A target of saliency 0.2 was used.

The model shows a difference in the ‘Eyes Open’ and ‘Eyes Closed’ conditions, with the first causing more gaze following. This effect is small at first, but

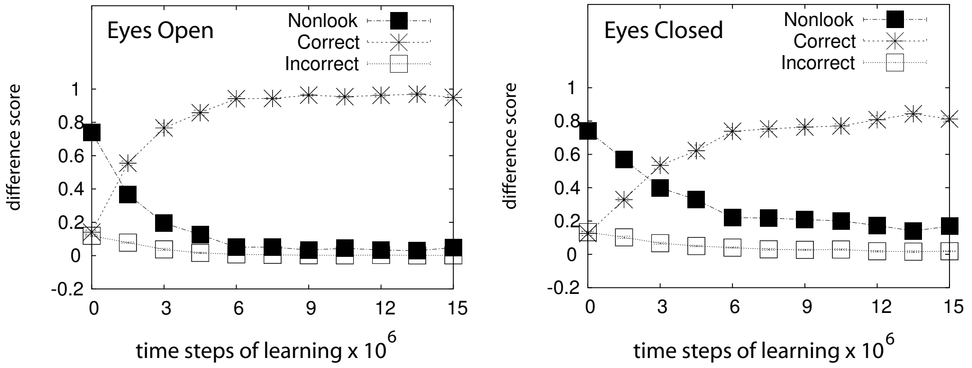


Figure V.5: Result details (nonlooks, correct, and incorrect responses) for simulation of Brooks & Meltzoff's conditions.

easier to measure with time. The reason for the different scores is that in the 'Eyes Open' condition both head and eye direction cues are present (both **h** and **e** add to the value of **m** in the direction of target), while in the 'Eyes Closed' condition only the head direction cue is present (i.e. all elements of **e** are set to zero because the eyes are closed), causing less turns. For reasons similar to those described above (Corkum & Moore's experimental simulations), it is possible that this difference might be present at earlier ages, but would not have been detected by Brooks and Meltzoff until about 10 months of age. More experimental repetitions, therefore, are encouraged.

V.C.3 Measuring looking time and checking behavior

Although not measured methodically, Tomasello (1995) reported that as infants progress in their gaze following skills, they tend to a) look longer at objects looked at by the experimenter, and b) alternate their gaze between experimenter and the object looked at more consistently. These observations are used to argue for a transition from attentional to mentalist gaze following. Tomasello took this as an indication of a transition to a mentalist gaze following.

To investigate if the model behaves like this, we again simulated Corkum & Moore's 'H + E' experiment using a target of saliency of 1. This time, trials were extended to 50 seconds, much longer than the standard 6 seconds, to make sure that the infant looked away from the experimenter at least once during the trial. Only trials where the infant looked from experimenter (at the beginning of the trial) to the target were considered. The number of time steps from the time the infant looks at the target to the time it looks away from it was measured. Fig. V.6 left shows the results. With time, the infant looks at the target longer, reaching a maximum early on, at around 400,000 time steps. The model explains this increase in time looking at the target as simply the infant doing more exploitation and less exploration (refer to softmax formula in chapter III), and not as taking a special interest because someone else is looking at it, as Tomasello suggested. Looking time reaches a maximum of around 2.2 when the model has fully learned the value of saliencies, and "exploits" the object's saliency, not looking away until habituation kicks in, making the other objects (including the experimenter) more attractive.

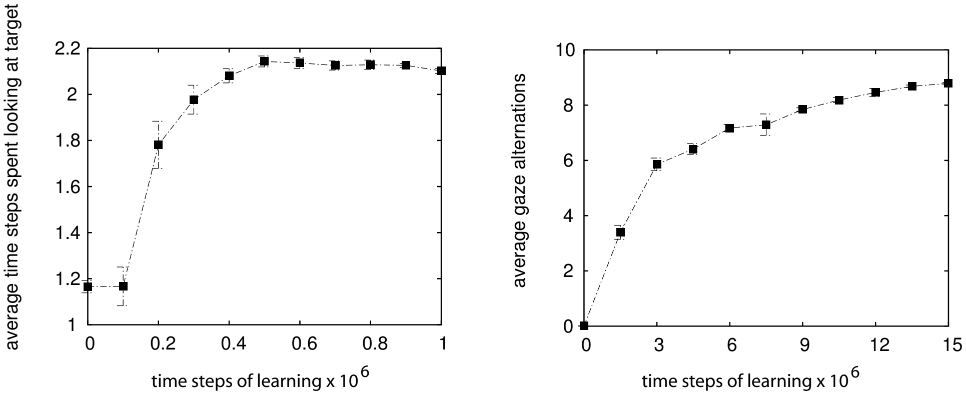


Figure V.6: Left: Average time looking at target. Right: Average number of gaze alternations. (Note the different scales used in the x-axis.)

Fig. V.6 right shows the average number of continuous gaze alternations between the experimenter and the target for the same experimental setup, until the cycle is broken (note the different time scale with respect to Fig. V.6 left). A gaze alternation consists of the infant model looking directly from the experimenter to the target and directly back, with no intermediate gazes in other directions. With time, the infant does more gaze alternations.

The model explains gaze alternations as stemming from the dynamics of habituation: When the infant model gets 'bored' looking at the experimenter, it looks in the direction of the object (because it is salient and because the experimenter is looking at it). As the infant model habituates to the object, it also dishabituates to the experimenter (because the experimenter is no longer in the infant's focus of attention), so that once it gets 'bored' looking at the object, it looks in the direction of the experimenter again. This process would repeat itself indefinitely, but since actions are chosen probabilistically, there is always a chance that the infant looks somewhere else, breaking the cycle. The number of alternations increases with time for a similar reason than above: the amount of exploration diminishes with time, in favor of exploitation, making it more difficult to break the cycle.

V.D Discussion

By replicating the experimental results showing the progressive incorporation of eye direction cues, the model shows an alternative account to the attentional/mentalist dichotomy: Instead, both head and eye direction cues are used to some degree throughout development, with eye direction cues being more important later in learning. The model thus unifies both attentional and mentalist accounts under a learning approach. It suggests that the incorporation of theory-of-mind, including gaze following, should be seen as happening along a continuum within a learning framework, where the eye direction cues are gradually incorporated in action selection; a similar view is proposed in Tomasello, Call, and Hare (2003).

V.E Acknowledgements

The text of this chapter, in part, is a reprint of the material in:

H. Jasso, J. Triesch, and G. O. Deák. Using eye direction cues for gaze following - a developmental model. In *Proceedings of the Fifth International Conference on Development and Learning (ICDL'06)*, Bloomington, IN, May 2006.

The dissertation author was the primary investigator and author of this paper.

VI

Simulations of Autism, and a Link to the Mirror Neuron System

Autism is a neurodevelopmental disorder characterized by delays in social responsiveness and language as used in social communication, as well as repetitive behaviors (Dawson, Toth, Abbott, Osterling, Munson, Estes, and Liaw, 2004). An early indicator of autism is a significant delay in the development of joint attention behaviors, including gaze following, with respect to control groups (Dawson, Meltzoff, Osterling, Rinaldi, and Brown, 1998; Loveland and Landry, 1986; Mundy, Sigman, and Kasari, 1990; Pelphrey, Morris, and McCarthy, 2005). This chapter focuses on possible reasons for this delay.

The first section of this chapter shows how the model supports an explanation based on an aversion to social stimuli such as faces and eye contact. The second section explores the relation between this aversion, the development of mirror neuron systems (Rizzolatti and Craighero, 2004), and autism, within the context of the gaze following model.

VI.A The Effect of Social Stimuli Aversion in the Development of Gaze Following

Autistic individuals tend to dislike social stimuli, and in particular faces (Adrien et al., 1993; Chawarska et al., 2003; Maestro et al., 2002; Tantam et al., 1993; Klin et al., 2003; Dawson et al., 1998) and eye contact (Hutt and Ounsted, 1966). To test if this could delay the development of gaze following, the model was run using different values of the caregiver's saliency (Φ_C) (as subjectively perceived by the infant), ranging from negative 1 (aversive) to 4 (attractive). The basic gaze following experiment described in chapter III was performed. Fig. VI.1 shows how basic gaze following performance in the model decreases as the perceived caregiver's saliency decreases, showing no gaze following for zero or negative caregiver saliencies.

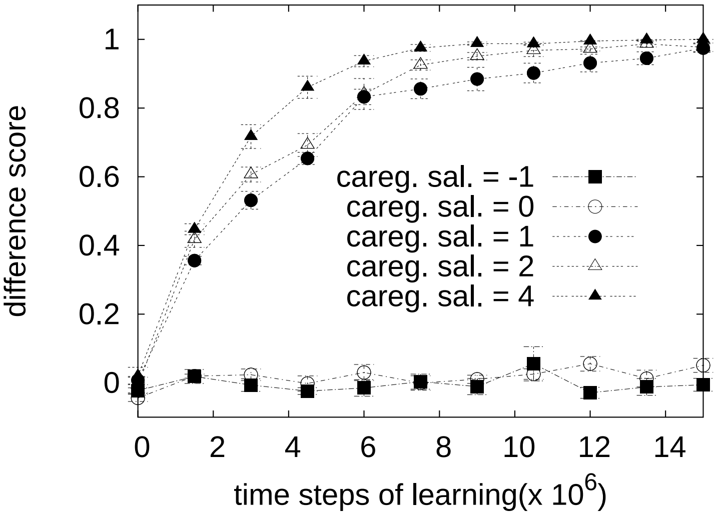


Figure VI.1: Gaze following performance for different caregiver saliencies (Φ_C). For low caregiver saliency gaze following will emerge only very slowly or not at all. Error bars represent standard error of the mean (5 repetitions).

The explanation given by the model of why a diminished perceived caregiver saliency could affect gaze following performance is that the infant will tend to disregard or avoid the caregiver, losing opportunities to learn the correlation between head/eyes direction and the location of objects in the room. Fig. VI.2 shows how indeed the infant avoids the caregiver in those cases: it shows the percentage time steps that the infant looks at the caregiver during a freeplay period of 100,000 time steps conducted after all training is finished (i.e. after 15,000,000 time steps of learning). The dynamics of the environment, and the behavior of the caregiver are the same during this freeplay period than in typical learning period. When the perceived caregiver saliency is zero or negative, the model infant spends practically no time looking at the caregiver.

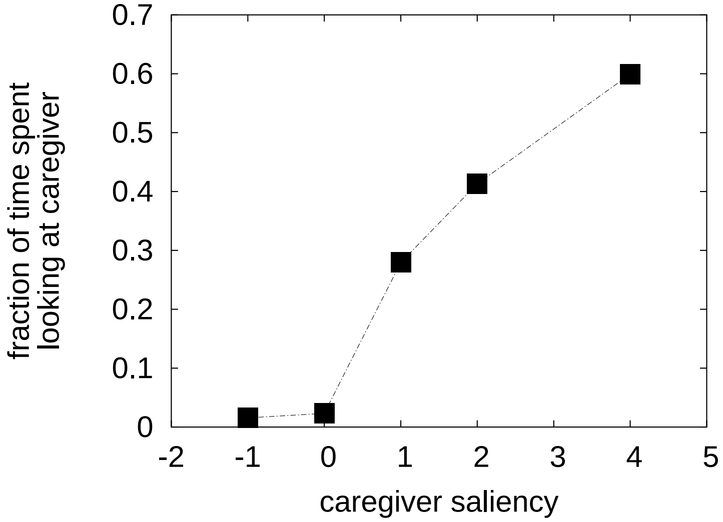


Figure VI.2: Fraction of time looking at caregiver during freeplay for different caregiver saliencies.

VI.B Gaze Following, Mirror Neurons, and Autism

Mirror neurons are a class of pre-motor neurons originally found in macaque area F5, and likely to exist in humans too (Rizzolatti and Craighero, 2004). Their defining characteristic is that they become activated when the animal performs an action such as reaching for a fruit and grasping it or when the animal sees another agent perform the same or a similar action. Because of this property, mirror neurons are believed to play a role in a number of social-cognitive functions (Rizzolatti, 2005) including understanding other's actions, imitation, and intention understanding (Pobric and Hamilton, 2006; Iacobini, Woods, Brass, Bekkering, Mazziotta, and Rizzolatti, 1999; Iacobini, Molnar-Szakacs, Gallese, Buccino, Mazziotta, and Rizzolatti, 2005). This section investigates the possible characterization of a component of the model as a mirror neuron system. The rationale for this is twofold: First, gaze following can be thought of as imitative behavior (the individual observes the other looking at an object, and imitates by looking at the same object (Nagai, 2005a,b; Hoffman, Grimes, Shon, and Rao, 2006)), and mirror neurons are thought to play a role in imitation (Iacobini et al., 1999). Second, autism has been associated with deficits in the mirror system (Daprett et al., 2006), and gaze following is diminished in autism (Pelphrey, Morris, and McCarthy, 2005).

VI.B.1 Mirror neuron system properties of the model

In the model, the **m** layer learns to behave like a mirror neuron system: Units in this layer become active (i.e. with positive values) when the infant looks at the location in space associated with them, and inactive (i.e. with values around zero) when then infant does not look there. This is illustrated by measuring the activation of two units in **m**: one corresponding to a location within the infant's field of view when looking at the caregiver (neuron A in Fig. VI.3) and another corresponding to a location outside the infant's field of view (neuron B in Fig. VI.3). These units are measured during the 100,000 time step interval described above, where the infant interacts with the caregiver in freeplay. Fig. VI.4 shows the distributions of activations of the two neurons when the infant shifts gaze to this location vs. when the infant shifts gaze to a different location (to the opposite side of the room). This is calculated for both units, for situations when there is an object in that location, and when the location is empty. (Note that because of the softmax formula for action selection (see chapter III for a description of the softmax formula), a unit could have negative values and still be selected.) These units can be viewed as pre-motor neurons because their activation increases the probability of performing the associated action, without automatically leading to the corresponding gaze shift. Instead, multiple such action plans will usually compete for being executed. In addition, execution of any action may be inhibited by additional brain structures which are not included in the model.

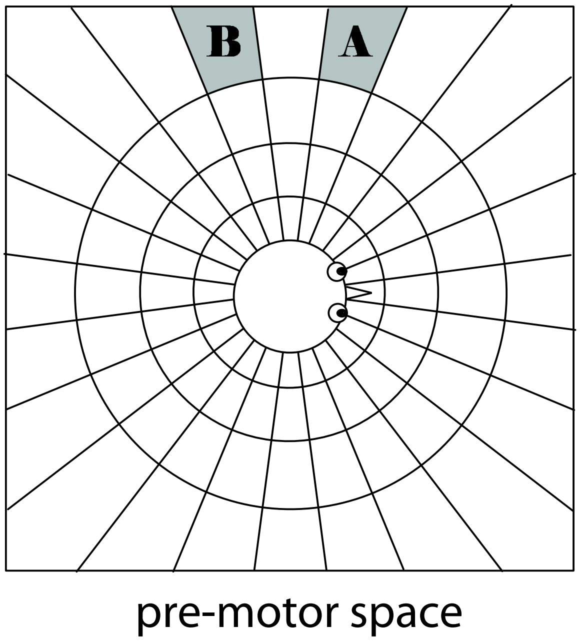


Figure VI.3: Location of pre-motor units whose motor and sensor properties were measured (see Fig. VI.4 and Fig. VI.5).

Similarly, Fig. VI.5 shows how units in this layer become active when the caregiver looks in the direction of their target location, and become inactive when the caregiver does not look in that direction. Measurements are done after training, in an experimental setup where the infant and caregiver sit facing each other, and the caregiver looks at either towards a location corresponding to unit A/B's action, or towards a location on the opposite side of the room. The figure shows that when the caregiver looks in the direction corresponding to a unit's action, the activation corresponding to that unit increases.

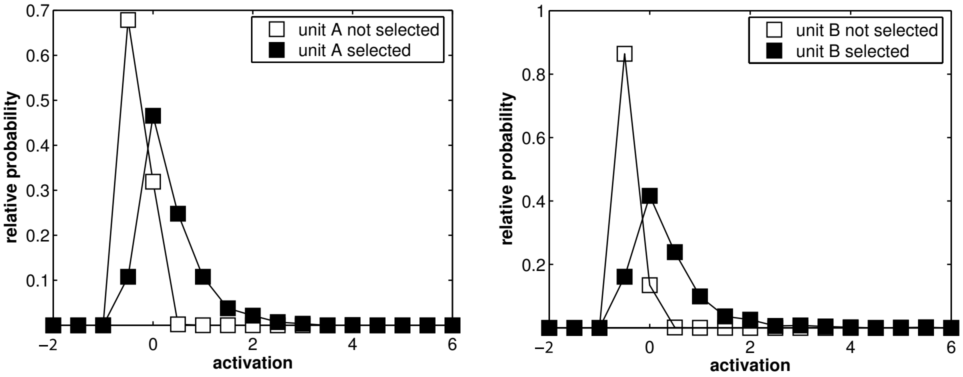


Figure VI.4: Illustration of motor properties of two pre-motor units in layer **m** (see Fig. VI.3 for location of units). See text for details. The sensor properties of the same two units are illustrated in Fig. VI.5.

Since these units become active for doing or observing action, they fulfill the defining properties of mirror neurons.

VI.B.2 Effect of social stimuli aversion in 'neural mass'

The activations of the units in **m** are defined by elements of the **M** weight matrix. For gaze following, this corresponds to the **h** and **e** sections of **M** (see Fig. III.5). Fig. VI.6 left shows how the sum of the absolute values of these weights (which can be thought of as 'neural mass') is reduced with the aversion to social stimuli explored in the previous section. The model, therefore, supports the idea that there is a link between the development of the mirror neuron system and autism. It gives a possible explanation to this, as the result of an aversion to social stimuli.

VI.C Discussion

Besides showing a possible link between aversion to social stimuli and autism, the model gives some insights on the possible developmental origins of mirror neurons. The idea of a learned mirror neuron system has been proposed

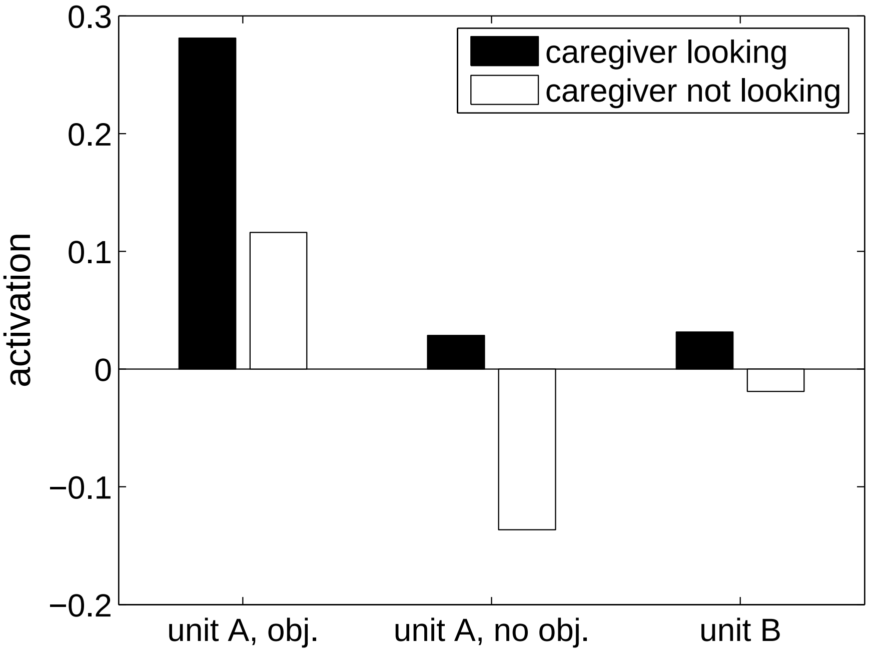


Figure VI.5: Illustration of sensor properties of two pre-motor units in layer **m** (see Fig. VI.3 for location of units). See text for details. The motor properties of the same two units are illustrated in Fig. VI.4.

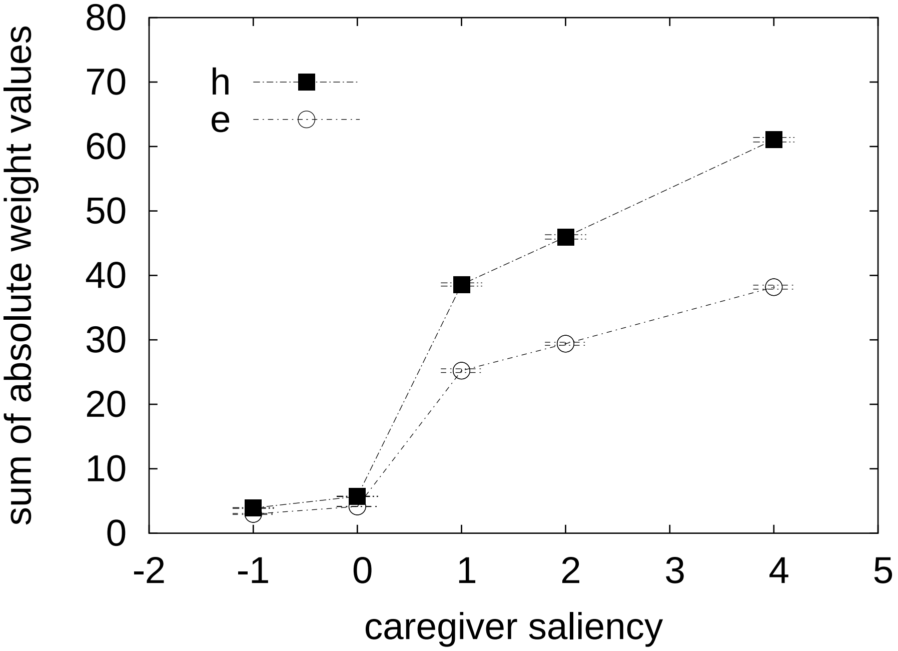


Figure VI.6: Plot of caregiver saliency vs. 'neural mass'. Error bars represent standard error of the mean (5 simulations).

before (Heyes, 2001; Oztop and Arbib, 2002; Keysers and Perrett, 2004; Brass and Heyes, 2005; Weber, Wermter, and Elshaw, 2001; Jones, 2006; Metta, Sandini, Natale, Craighero, and Fadiga, 2006; Oztop, Kawato, and Arbib, 2006), although these accounts are based on Hebbian learning in the context of self-observation and/or being imitated by other agents, not reinforcement learning. Finally, the model also supports the idea of using reinforcement learning as the learning method for other mirror neurons. This is discussed in more detail in (Triesch, Jasso, & Deák, in press) .

VI.D Acknowledgements

The text of this chapter, in part, is a reprint of the material in:

J. Triesch, H. Jasso, and G. O. Deák. Emergence of mirror neurons in a model of gaze following. In *Proceedings of the Fifth International Conference on*

Development and Learning (ICDL'06), Bloomington, IN, May 2006. The dissertation author was co-author of this paper.

and

J. Triesch, H. Jasso, and G. O. Deák. Emergence of Mirror Neurons in a Model of Gaze Following. *Adaptive Behavior*, (in press). The dissertation author was co-author of this paper.

VII

Gaze Following in a Virtual Reality Environment

The model presented in chapter III simulates the environment the infant model interacts with. The benefit over the use of physical robotic models is that of significantly reducing learning times and lowering model building costs. However, this leads to some assumptions being made, such as ignoring the partial or total visual occlusion of objects, the reduced visual saliency of objects with distance, and the variable times needed for the completion of different actions. The use of robots is still desirable, so that the validity of the results and predictions of the model can be demonstrated for simplifications of the environment and the model.

A middle point between abstract computational models and physical robots can be found in the use of virtual reality simulations. These can be used to semi-realistically recreate three-dimensional aspects of the environment the infant model interacts with, with a fraction of the costs and development times associated with them. With this in mind, a *virtual reality platform* was built. Similar virtual reality models have been built to model non-verbal communication, such as how body language from a language teacher provides useful cues to the language learner to infer the referent of unknown words (Zhang, Yu, and Smith, 2006; Yu, Zhang, and Smith, 2006).

The virtual reality platform was designed to be flexible enough to build and test different cognitive development models. Such models, which are often based on connectionist (Elman et al., 1996) and dynamical systems (Thelen and Smith, 1994) approaches, stress the importance of interactions with the physical and social environment for cognitive development. Developmental schemes are also being proposed in the field of intelligent robotics (Asada et al., 2001; Brooks et al., 1998; Weng et al., 2001): instead of building a fully working robot, a body capable of interacting with the environment is given general learning mechanisms that allow it to evaluate the results of its actions. It is then "set free" in the world to learn a task through repeated interactions with both the environment and a human supervisor.

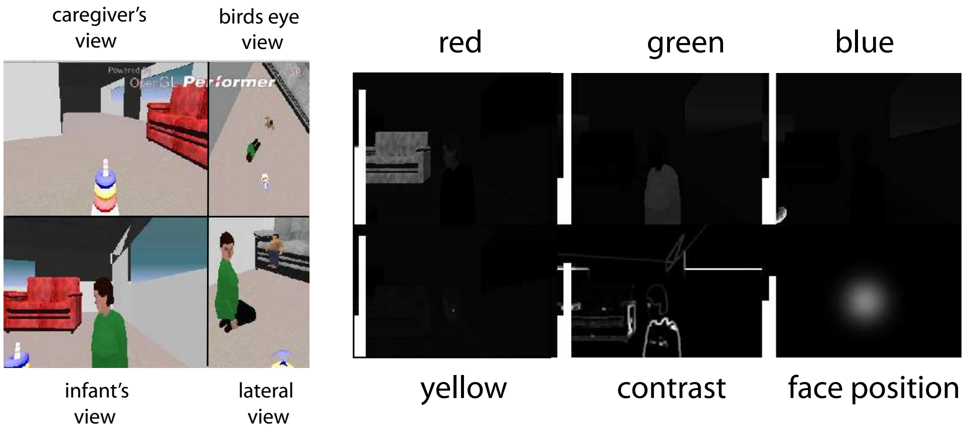


Figure VII.1: Left: various views of a virtual living room used to model the emergence of gaze following. Right: Saliency maps generated by analyzing the infant's visual input ("infant's view", left). White bars on left of each saliency map indicate the intrinsic reward of the associated feature and its current habituation level.

The first section of this chapter describes our modeling platform and the underlying software infrastructure. The second section shows how it is currently being used to build an embodied model of the emergence of gaze following in mother-infant interactions. The final section discusses the relative benefits of virtual vs. robotic modeling approaches.

VII.A The Platform

VII.A.1 Platform overview

The platform allows the construction of semi-realistic models of arbitrary visual environments. A virtual room with furniture and objects can be set up easily to model, say, a testing room used in a controlled developmental psychology experiment, or a typical living room. These visual environments are populated with virtual characters. The behavior and learning mechanisms of all characters can be specified. Typically, a virtual character will have a vision system that receives images from a virtual camera placed inside the character's head. The simulated vision system will process these images and the resulting representation will drive the character's behavior (Terzopoulos et al., 1994). Fig. VII.1 shows an example setting.

An overview of the software structure is given in Fig. VII.2. The central core of software, the "Simulation Environment," is responsible for simulating the learning agent (infant model) and its social and physical environment (caregiver model, objects, . . .). The Simulation Environment was programmed in C++ and will be described in more detail below. It interfaces with a number of 3rd party libraries for animating human characters (BDI DI-Guy), managing and rendering of the graphics (SGI OpenGL Performer), and visual processing of rendered images to simulate the agents' vision systems (OpenCV).

The platform currently runs on a Dell Dimension 4600 desktop computer with a Pentium 4 processor running at 2.8GHz. The operating system is Linux. An NVidia GeForce video graphics accelerator speeds up the graphical simulations.

VII.A.2 Third-party software libraries

OpenGL Performer. The Silicon Graphics *OpenGL Performer*[1] toolkit is used to create the graphical environment for running the experiments. OpenGL

[1] http://www.sgi.com/products/software/performer/

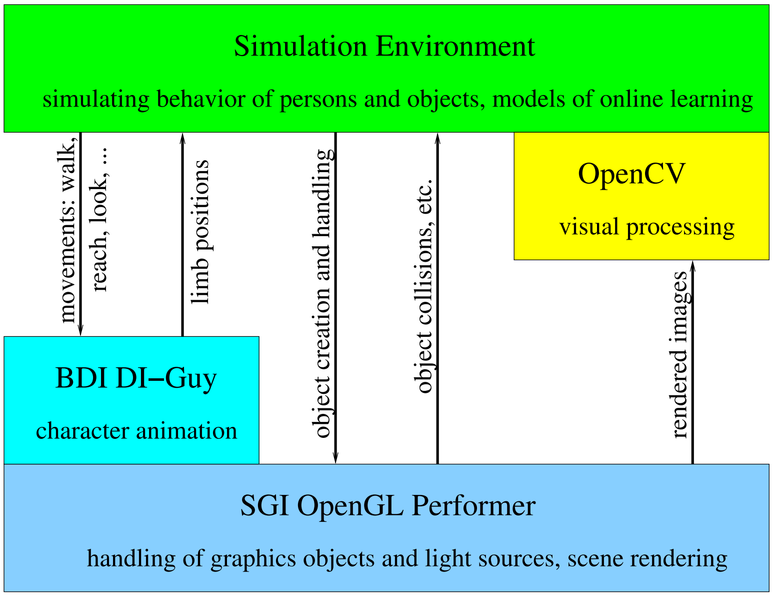

Simulation Environment
simulating behavior of persons and objects, models of online learning
OpenCV
visual processing
movements: walk, reach, look, ...
limb positions
object creation and handling
object collisions, etc.
rendered images
BDI DI–Guy
character animation
SGI OpenGL Performer
handling of graphics objects and light sources, scene rendering

Figure VII.2: Overview of software structure.

Performer is a programming interface built atop the industry standard *OpenGL* graphics library. It can import textured 3D objects in many formats, including OpenFlight (.flt extension) and 3D Studio Max (.3ds extension). OpenGL is a software interface for graphics hardware that allows the production of high-quality color images of 3D objects. It can be used to build geometric models, view them interactively in 3D, and perform operations like texture mapping and depth cueing. It can be used to manipulate lighting conditions, introduce fog, do motion blur, perform specular lighting, and other visual manipulations. It also provides virtual cameras that can be positioned at any location to view the simulated world.

DI-Guy. On top of OpenGL Performer, Boston Dynamics's *DI-Guy* libraries[2] provide lifelike human characters that can be created and readily inserted into the virtual world. They can be controlled using simple high-level commands such as "look at position (X, Y, Z)," or "reach for position (X, Y, Z) using the left arm," resulting in smooth and lifelike movements being generated automatically. The facial expression of characters can be queried and modified. DI-Guy

[2] `http://www.bdi.com`

provides access to the character's coordinates and link positions such as arm and leg segments, shoulders, hips, head, etc. More than 800 different functions for manipulating and querying the characters are available in all. Male and female characters of different ages are available, configurable with different appearances such as clothing style.

OpenCV. Querying the position of a character's head allows us to dynamically position a virtual camera at the same location, thus accessing the character's point of view. The images coming from the camera can be processed using Intel's *OpenCV* library[3] of optimized visual processing routines. OpenCV is an open-source, extendable software intended for real-time computer vision, and is useful for object tracking, segmentation, and recognition, face and gesture recognition, motion understanding, and mobile robotics. It provides routines for image processing such as contour processing, line and ellipse fitting, convex hull calculation, and calculation of various image statistics.

VII.A.3 The simulation environment

The Simulation Environment comprises a number of classes to facilitate the creation and running of simulations. Following is a description of the most important ones.

The Object Class. The Object class is used to create all inanimate objects (walls, furniture, toys, etc.) in the simulation. Instances of the Object class are created by giving the name of the file containing the description of a 3D geometrically modeled object, a name to be used as a handle, a boolean variable stating whether the object should be allowed to move, and its initial scale. The file must be of a format readable by OpenGL Performer, such as 3D Studio Max (.3ds files) or OpenFlight (.flt files). When an Object is created, it is attached to the Performer environment. There are methods for changing the position of the Object, for rotating it, and changing its scale. Thus, it can easily be modeled that

[3] `http://www.intel.com/research/mrl/research/opencv/`

characters in the simulation can grasp and manipulate objects, if this is desired.

The Object Manager Class. The Object Manager class holds an array of instances of the Object class. The Object Manager has methods for adding objects (which must be previously created) to the scene, removing them, and querying their visibility from a specific location. The latter function allows to assess if, e.g., an object is within the field of view of a character, or if the character is looking directly at an object.

The Person Class. The Person class is used to add any characters to the simulation. These may be rather complicated models of, say, a developing infant simulating its visual perception and learning processes, or they may be rather simplistic agents that behave according to simple scripts. To create an instance of the Person class, a DI-Guy character type must be specified, which determines the visual appearance of the person, along with a handle to the OpenGL Performer camera assigned to the character. The Brain type and Vision System type (see below) must be specified. If the character's actions will result from a script, then a filename with the script must be given. For example, such a script may specify what the character is looking at at any given time. One Brain object and one Vision System object are created, according to the parameters passed when creating the Person object. The Person object must be called periodically using the "update" method. This causes the link corresponding to the head of the character to be queried, and its coordinates to be passed to the virtual camera associated with the character. The image from the virtual camera in turn is passed to the character's Vision System, if the character has any. The output of the Vision System along with a handle to the DI-Guy character is passed to the Brain object, which will decide the next action to take and execute it in the DI-Guy character.

The Brain class. The Brain class specifies the actions to be taken by an instance of the Person class. The space of allowable actions is determined by the DI-Guy character type associated with the person. The simplest way of how a Brain object can control the actions of a Person is by following a script. In this case the Person will "play back" a pre-specified sequence of actions like a tape recorder. More interestingly, a Brain object can contain a simulation of the person's nervous system (at various levels of abstraction). The only constraint is that this simulation has to run in discrete time steps. For example, the Brain object may instantiate a reinforcement learning agent (Sutton, 1998) whose state information is derived from a perceptual process (see below) and whose action space is the space of allowable actions for this character. An "update" method is called every time step to do any perceptual processing, generate new actions, and possibly simulate experience dependent learning.

The actions used to control a character are fairly high-level commands such as "look to location (X,Y,Z)," "walk in direction Θ with speed v," or "reach for location (X,Y,Z) with the left arm," compared to direct specification of joint angles or torques. Thus, this simulation platform is not well suited for studying the development of such motor behaviors. Our focus is on the development of higher-level skills that use gaze shifts, reaches, etc. as building blocks. Thus, it is assumed that elementary behaviors such as looking and reaching have already developed and can be executed reliably in the age group of infants being modeled — an assumption that of course needs to be verified for the particular skills and ages under consideration. The positive aspect of this is that it allows to focus efforts on modeling the development of higher level cognitive processes without having to worry about such lower-level skills. This is in sharp contrast to robotic models of infant development, where invariably a significant portion of time is spent on implementing such lower level skills. In fact, skills like two-legged walking and running, or reaching and grasping are still full-blown research topics in their own right in the area of humanoid robotics.

The Vision System class. The Vision System class specifies the processing to be done on the raw image corresponding to the person's point of view (as extracted from a virtual camera dynamically positioned inside the person's head). It is used to construct a representation of the visual scene that a Brain object can use to generate behavior. Thus, it will typically contain various computer vision algorithms and/or some more specific models of visual processing in human infants, depending on the primary goal of the model.

If desirable, the Vision System class may also use so-called "oracle vision" to speed up the simulation. Since the simulation environment provides perfect knowledge about the state of all objects and characters in the simulation, it is sometimes neither necessary nor desirable to infer such knowledge from the rendered images through computer vision techniques, which can be difficult and time consuming. Instead, some property, say the identity of an object in the field of view, can simply be looked up in the internal representations maintained by the simulation environment — it functions as an oracle. This simplification is desirable if the visual processing (in this case object recognition) is not central to the developmental process under consideration, and if it can be assumed that it is sufficiently well developed prior to the developmental process being studied primarily. In contrast, in a robotic model of infant development, there is no "oracle" available, which means that all perceptual processes required for the cognitive skill under consideration have to be modeled explicitly. This is time-consuming and difficult.

Main Program and Control Flow. The main program is written in C++ using object-oriented programming. OpenGL Performer is first initialized, and a scene with a light source is created and positioned. A window to display the 3D world is initialized, and positioned on the screen. Virtual cameras are created and positioned in the world, for example as a birds eye view or a lateral view. Cameras corresponding to the characters are created but positioned dynamically as the characters move their heads. Each camera's field of view can be set (characters would usually have around a 90^o field of view), and can be configured to eliminate objects that are too close or too far. All cameras created are linked to the window that displays the 3D world. Environment settings such as fog, clouds, etc. can be specified. The DI-Guy platform is then initialized, and a scenario is created. The scenario holds information about all the characters, and must be used to create new characters. New instances of the Person class are created, and their activities are specified by periodically giving them new actions to perform. The level of graphical detail of the characters can be specified to either get fairly realistically looking characters or to speed up processing.

Statistics gathering. Throughout the session, statistics are gathered by querying the different libraries: DI-Guy calls can be used to extract the position of the different characters or the configuration of their joints. The Object Manager can be used to query the position of objects and their visibility from the point of view of the different characters. In addition, the internal states of all characters' simulated nervous systems are perfectly known. This data or arbitrary subsets of it can easily be recorded on a frame by frame basis for later analysis. These statistics are useful for analyzing long-term runs, and allow to evaluate whether the desired behavior is being achieved and at what rate. We point out that every simulation is perfectly reproducible and can be re-run if additional statistics need to be collected.

VII.B A Virtual Reality Environment for Modeling Gaze Following

This section describes how the virtual platform is being used to implement a realistic version of the model of gaze following presented in this thesis.

The platform was configured for an experimental setup consisting of a living room with furniture and objects, all of them instantiations of the Object class and built from 3D Studio Max objects. Two instantiations of the Person class are created, one for the caregiver and one for the baby. The caregiver and learning infant are placed facing each other. The caregiver instantiates a Brain object controlling its behavior. The positions of objects are fed to the caregiver's

Brain. No visual system is given to the caregiver.

The baby instantiates a Visual System object that models a simple infant vision system. In particular, it evaluates the *saliency* of different portions of the visual field (Itti and Koch, 2000), it recognizes the caregiver's head, and it discriminates different head poses of the caregiver. Saliency computation is based on six different features, each habituating individually according to Stanley's model of habituation (Stanley, 1976). The feature maps (see Fig.VII.1) are: red, green, blue and yellow color features based on a color opponency scheme (Lee et al., 2002), a contrast feature that acts as an edge detector by giving a high saliency to locations in the image where the intensity gradient is high, and finally a face detector feature that assigns a high saliency to the region of the caregiver's face, which is localized through oracle vision. The saliency of the face can be varied depending on the pose of the caregiver's face with respect to the infant (infant sees frontal view vs. profile view of the caregiver). A similar scheme for visual saliency computation has been used in (Breazeal, 2002) for a non-developing model of gaze following, using skin tone, color, and motion features.

A reinforcement learning system such as the one described in chapter 3 can be incorporated into the infant's Brain object. The infant should learn to direct gaze to the caregiver to maximize visual reward, and habituation will cause it to look elsewhere before looking back to the caregiver. With time, the infant should learn to follow the caregiver's line of regard.

VII.B.1 Platform performance

To illustrate the performance of the platform given the current hardware, a number of measurements were made to establish the computational bottlenecks for this specific model. The time spent for each frame was divided into three separate measures for analysis: the time to calculate the feature maps (Vision), the time to display them (Map Display), and the time for the DI-Guy environment to calculate the next character positions and display them (Animation). Table VII.1

shows how the times vary with the resolution of the infant's vision system. As can be seen, most time is spent on simulating the infant's visual processing. Real time performance is achievable if the image resolution is not set too high.

Table VII.1: Simulation times per frame (in seconds).

Image Scale	**Vision**	**Map Display**	**Animation**
80×60	0.0226	0.0073	0.0476
160×120	0.0539	0.0092	0.0431
240×180	0.0980	0.0121	0.0522
320×240	0.1507	0.0113	0.0422
400×300	0.2257	0.0208	0.0507
480×360	0.3025	0.0276	0.0539

Table VII.2: Robotic vs. virtual models of infant cognitive development.

Property	**Robotic Model**	**Virtual Model**
physics	real	simplified or ignored
agent body	difficult to create	much easier to simulate
motor control	full motor control problem	substantially simplified
visual environment	realistic	simplified computer graphics
visual processing	full vision problem	can be simplified through oracle vision
social environment	real humans	real humans or simulated agents
real time requirements	yes	no, simulation can be slowed down or sped up
data collection	difficult	perfect knowledge of system state
reproducibility of experiments	difficult	perfect
ease-of-use	very difficult	easy
development costs	extremely high	very modest

VII.C Discussion

The platform presented here is particularly useful for modeling the development of *embodied* cognitive skills. In the case of the emergence of gaze following discussed above, it is suitable because the skill is about the inference of mental states from bodily configurations, such as head and eye position, which are realistically simulated in our platform.

VII.C.1 Virtual vs. robotic models

Recently, there has been a surge of interest in building robotic models of cognitive development. Compared to the virtual modeling platform presented here, there are a number of important advantages and serious disadvantages of robotic models that we will discuss in the following. A summary of this discussion is given in Table VII.2.

Physics. The virtual simulation is only an approximation of real-world physics. The movements of the characters do not necessarily obey physical laws but are merely animated to “look realistic.” For the inanimate objects, we currently do not simulate any physics at all. In a robotic model, the physics are real, of course. The justification of neglecting physics in the virtual model is that the cognitive skills we are most interested in are fairly high-level skills, i.e., we simply do not want to study behavior at the level of muscle activations, joint torques, and frictional forces, but at the level of primitive actions such as gaze shifts, reaches, etc., and their coordination into useful behaviors.

Agent body. In the virtual modeling platform, we can choose from a set of existing bodies for the agents. These bodies have a high number of degrees of freedom, comparable to that of the most advanced humanoid robots. Further, since physics is not an issue, we are not restricted by current limitations in robotic actuator technology. Our characters will readily run, crawl, and do many other things.

Motor control. Our interface to the agents in the model allows us to specify high-level commands (walk here, reach for that point, look at this object). The underlying motor control problems do not have to be addressed. In contrast, for a robotic model the full motor control problem needs to be solved, which represents a major challenge. Clearly, the platform should not be used to study the specifics of human motor control but it makes it much easier to focus on higher level skills. At the same time, perfect control over individual joint angles is possible, if desired.

Visual environment. The simulated computer graphics environment is of course vastly simpler than images taken by a robot in a real environment. For example, shadows and reflections are not rendered accurately, and the virtual characters are only coarse approximations of human appearance. Clearly, again, such a modeling platform should not be used to, say, study the specifics of human object recognition under lighting changes. The skills we are most interested in, however, use object recognition as a basic building block (e.g., the ability to distinguish different head poses of the caregiver with a certain accuracy). We believe that the details of the underlying mechanism are not crucial as long as the level of competence is accurately captured by the model.

Visual processing. In the virtual modeling platform we can vastly simplify perceptual processes through the use of oracle vision. In a robotic model, this is not possible and the perceptual capabilities required for some higher level cognitive skills may simply not have been achieved by contemporary computer vision methods. In this situation, it is common practice in robotics to drastically simplify the environment and objects such that simple vision methods become sufficient.

Social environment. A robotic model can interact with a real social environment, i.e., one composed of real human beings. In our virtual modeling platform we could achieve this to some extent by using standard Virtual Reality interfaces such as head mounted displays in conjunction with motion tracking devices. In such a setup a real person would control a virtual person in the simulation, seeing what the virtual person is seeing through the head mounted display. However, the ability to experiment with vastly simplified agents as the social environment allows us to systematically study what aspects of the social environment, i.e., which behaviors of caregivers, are really crucial for the development of specific social skills (Teuscher and Triesch, 2004). This degree of control over the social environment cannot be achieved with human subjects. Also, the social agents may be programmed to exhibit behavior that replicates important statistics of caregiver behavior observed in real infant caregiver interactions. For example, Deák et al. are collecting such statistics from videos of infant-caregiver dyad interactions (Deák, Wakabayashi, and Jasso, 2004). We are planning on developing caregiver models that closely replicate the observed behaviors.

Real time requirements. A robotic model must be able to operate in real time. This severely limits the complexity of the model. Perceptual processes in particular are notoriously time consuming to simulate. In the virtual model, we are not restricted to simulating in real time. Simulations may be slowed down or sped up arbitrarily. In addition, the availability of oracle vision allows to save precious computational resources.

Data collection. In the virtual model it is trivial to record data about every smallest detail of the model at any time. This is much harder to achieve in a robotic model interacting with real human caregivers. In particular, the exact behavior of the caregiver is inherently difficult to capture. Useful information about the caregiver behavior can be recovered by manually coding video records of the experiment, but this information is not available at the time of the experiment.

Reproducibility of experiments. Along similar lines, the virtual modeling platform allows perfect reproducibility of experiments. Every last pixel of the visual input to the learning agent can be recreated with fidelity. This is simply impossible in a robotic model.

Ease-of-use. Not having to deal with robotic hardware shortens development times, reduces maintenance efforts to a minimum, and makes it much easier to exchange model components with other researchers. Also, recreating the specific setup of a real-world behavioral experiment only requires changing a configuration file specifying where walls and objects are, rather than prompting a renovation.

Development costs. Finally, robotic models are much more expensive.

Most of the software components used in our platform (Linux OS, SGI OpenGL Performer, Intel OpenCV) are freely available to researchers. The lion share of the costs is the price of the BDI DI-Guy software.

All these benefits may make a virtual model the methodology of choice. Even if a robotic model is ultimately desirable, a virtual model may be used for rapid proto-typing. We see the use of virtual and robotic models as complementary. In fact, we are pursuing both methodologies at the same time in our lab (Kim et al., 2004).

VII.C.2 Possible extensions

There are several extensions to our platform that may be worth pursuing. First, we have only considered monocular vision. It is easy to incorporate binocular vision by simply placing two virtual cameras side by side inside a character's head. Foveation could also be added to the characters' vision systems. Second, in order to model language acquisition, a simulation of vocal systems and auditory systems of the characters could be added. Even in the context of non-verbal communication, a caregiver turning his head to identify the source of a noise may be a powerful training stimulus for the developing infant. Third, the platform is not restricted to modeling human development, but could be extended to model, say, the development of cognitive skills in a variety of non-human primates. To this end the appropriate graphical characters and their atomic behaviors would have to be designed. Fourth, on the technical side, it may be worth investigating in how far the simulation could be parallelized to run on a cluster of computers.

VII.D Acknowledgements

The text of this chapter, in part, is a reprint of the material in:

H. Jasso and J. Triesch. A virtual reality platform for modeling cognitive development. In *Proceedings of the Third International Conference on Develop-*

ment and Learning (ICDL'04), La Jolla, CA, October 2004.

The dissertation author was the primary investigator and author of this paper.

VIII

Conclusion

The importance of gaze following as a cornerstone skill for the infant's integration into the adult world is well established. However, the exact nature of the mechanism behind gaze following is still not well known. This dissertation presents a new computational model of gaze following built expressly to replicate its developmental trajectory as measured in key experimental observations. In doing so, the model shows how attentional and mentalist interpretations of gaze following can be unified into a learning account. It also offers a parsimonious, single-mechanism account of the improvement of spatial aspects of gaze following. This is done using a precise and reproducible, biologically plausible reinforcement learning algorithm. What the model proposes, therefore, is a mechanism where other's visual attention is not represented explicitly, but instead is implicitly encoded within a learning algorithm whose goal is to maximize rewards. Transitions in the gaze following abilities are similarly based on a progressive refinement of a reward-maximizing strategy based on experience. This approach to modeling based on reward maximization can be used to explore other social understanding themes such as social referencing (Feinman, 1982) and theory-of-mind (Premack and Woodruff, 1978). Hopefully, it will help clarify notions such as what it means to understand the attentional state, beliefs, and intentions of others, all central concepts in the study of social understanding.

IX

Appendix

Table IX.1: Results for Butteworth's two-target setting, from Butterworth and Cochran (1980) and Butterworth and Jarrett (1991). "Total number of trials" adds trials with Correct, Wrong, and Non-Codable responses, plus those where infants turned to the wrong side of the room. No measurements for 30°, 60°, or 90° were made at 18 months of age because of near-perfect scores at 6 and 12 months.

Age (months)	Experimental Setting	Total number of trials (2 trials) (per subject)	Correct	Wrong	Non-Codable Response
6	30°	24	11	0	9
6	60°	24	12	0	9
6	90°	24	14	0	7
6	120°	24	7	2	10
6	150°	24	2	5	13
12	30°	36	31	1	3
12	60°	36	32	0	4
12	90°	36	25	2	8
12	120°	36	8	18	9
12	150°	36	3	16	15
18	120°	36	16	3	20
18	150°	36	19	5	14

Table IX.2: Results for Butteworth's four-target setting, from Butterworth and Jarrett (1991). "Total number of trials" adds trials with Correct, Wrong, and Non-Codable responses, plus those where infants turned to the wrong side of the room.

Age (months)	Target Angle	Distracter Angle	Target Position Along Scan Path	Total number of trials (2 trials per subject)	Correct	Wrong	Non-Codable Response
6	30°	90°	1st	36	22	1	9
6	60°	120°	1st	36	18	1	11
6	90°	150°	1st	36	9	2	21
6	90°	30°	2nd	36	7	9	12
6	120°	60°	2nd	36	2	11	19
6	150°	90°	2nd	36	1	8	22
12	30°	90°	1st	36	26	0	9
12	60°	120°	1st	36	29	0	4
12	90°	150°	1st	36	21	3	11
12	90°	30°	2nd	36	16	8	9
12	120°	60°	2nd	36	12	14	7
12	150°	90°	2nd	36	6	15	11
18	30°	90°	1st	36	30	0	6
18	60°	120°	1st	36	27	0	8
18	90°	150°	1st	36	25	5	5
18	90°	30°	2nd	36	22	2	7
18	120°	60°	2nd	36	10	19	5
18	150°	90°	2nd	36	8	17	9

Bibliography

L. B. Adamson. *Communication Development During Infancy.* Westview Press, Boulder, CO, 1995.

J. L. Adrien, P. Lenoir, J. Martineau, and A. Perrot. Blind ratings of early symptoms of autism based upon family home movies. *Journal of the American Academy of Child and Adolescent Psychiatry*, 32:617–626, 1993.

S. M. Anstis, J. W. Mayhew, and T. Morley. The perception of where a television portrait is looking. *American Journal of Psychology*, 82:472–489, 1969.

M. Asada, K. F. MacDorman, H. Ishiguro, and Y. Kuniyoshi. Cognitive developmental robotics as a new paradigm for the design of humanoid robots. *Robotics and Autonomous Systems*, 37:185–193, 2001.

J. W. Astington, editor. *The Child's Discovery of the Mind.* Harvard University Press, Cambridge, MA, 1993.

J. W. Astington, P. L. Harris, and D. R. Olson, editors. *Developing Theories of Mind.* Cambridge University Press, Cambridge, MA, 1988.

S. Baron-Cohen. Precursors to a theory of mind: understanding attention in others. In A. Whiten, editor, *Natural Theories of Mind: Evolution, Development, and Simulation of Everyday Mindreading*, pages 233–251. Basil Blackwell, Cambridge, MA, 1991.

S. Baron-Cohen. The eye direction detector (EDD) and the shared attention mechanism (SAM): Two cases for evolutionary psychology. In C. Moore and P. J. Dunham, editors, *Joint Attention: Its Origins and Role in Development*, pages 41–59. Erlbaum, Hillsdale, NJ, 1995a.

S. Baron-Cohen. *Mindblindness: An Essay on Autism and Theory of Mind.* MIT Press, Cambridge, MA, 1995b.

S. Baron-Cohen, H. Tager-Flusberg, and D. J. Cohen, editors. *Understanding Other Minds: Perspectives From Autism.* Oxford University Press, Oxford, 2000.

J. Barresi and C. Moore. Intentional relations and social understanding. *Behavioral and Brain Sciences*, 19:107–154, 1996.

C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, Oxford, 1995.

M. Brass and C. Heyes. Imitation: Is cognitive neuroscience solving the correspondence problem? *Trends in Cognitive Sciences*, 9:489–495, 2005.

T. B Brazelton, B. Koslowski, and M. Main. The origin of reciprocity: The early mother-infant interaction. In M. Lewis and L. A. Rosenblum, editors, *The Effect of the Infant On Its Caretaker*, pages 49–76. Wiley, New York, 1974.

C. L. Breazeal. *Designing Sociable Robots.* MIT Press, Cambridge, MA, 2002.

I. Bretherton. Intentional communication and the development of mind. In D. Frye and C. Moore, editors, *Children's Theories of Mind: Mental States and Social Understanding*, pages 446–462. Erlbaum, Hillsdale, NJ, 1991.

G. W. Bronson. Changes in infants' visual scanning across the 2- to 14-week age period. *Journal of Experimental Child Psychology*, 49:101–125, 1990.

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

R. Brooks and A. N. Meltzoff. The development of gaze following and its relation to language. *Developmental Science*, 8:535–543, 2005.

R. A. Brooks, C. Breazeal, R. Irie, C. C. Kemp, M. Marjanovic, B. Scassellati, and M. M. Williamson. Alternative essences of intelligence. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, Madison, WI, July 1998.

R. A. Brooks, C. Breazeal, M. Marjanovic, B. Scassellati, and M. M. Williamson. The COG project: Building a humanoid robot. In C. L. Nehaniv, editor, *Computation for Metaphors, Analogy, and Agents*, volume 1562 of Springer Lecture Notes in Artificial Intelligence, pages 52–87. Springer-Verlag, 1999.

J. S. Bruner. *Child's Talk: Learning to Use Language.* Norton, New York, 1983.

J. S. Bruner. From joint attention to the meeting of minds: An introduction. In C. Moore and P. J. Dunham, editors, *Joint Attention: Its Origins and Role in Development*, pages 1–14. Erlbaum, Hillsdale, NJ, 1995.

G. Butterworth. The ontogeny and phylogeny of joint visual attention. In A. Whiten, editor, *Natural Theories of the Mind: The Evolution, Development and Simulation of Second-Order Mental Representations*, pages 223–232. Blackwell, Oxford, 1991.

G. Butterworth. Origins of mind in perception and action. In C. Moore and P. J. Dunham, editors, *Joint Attention: Its Origins and Role in Development*, pages 29–40. Erlbaum, Hillsdale, NJ, 1995.

G. Butterworth and E. Cochran. Towards a mechanism of joint visual attention in human infancy. *International Journal of Behavioral Development*, 3:253–272, 1980.

G. Butterworth and L. Grover. The origins of referential communication in human infancy. In L. Weiskrantz, editor, *Thought Without Language*, pages 5–24. Oxford University Press, Oxford, 1988.

G. Butterworth and L. Grover. Joint visual attention, manual pointing, and preverbal communication in human infancy. In M. Jeannerod, editor, *Attention and Performance XIII*, pages 605–624. Erlbaum, Hillsdale, NJ, 1990.

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

R. W. Byrne and A. Whiten, editors. *Machiavellian Intelligence. Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans.* Oxford University Press, Oxford, 1988.

E. Carlson and J. Triesch. A computational model of gaze following. In *Proceedings of the Eighth Neural Computation and Psychology Workshop, Connectionist Models of Cognition and Perception II*, Canterbury, UK, August 2003.

A. J. Caron, S. C. Butler, and R. Brooks. Gaze following at 12 and 14 months: Do the eyes matter? *British Journal of Developmental Psychology*, 20:225–239, 2002.

M. Carpenter, K. Nagell, M. Tomasello, G. Butterworth, and C. Moore. *Social Cognition, Joint Attention, and Communicative Competence From 9 to 15 Months of Age (Monographs of the Society for Research in Child Development).* University of Chicago Press, Chicago, 1998.

M. Carpenter, M. Tomasello, and S. Savage-Rumbaugh. Joint attention and imitative learning in children, chimpanzees, and enculturated chimpanzees. *Social Development*, 4:217–237, 1995.

K. Chawarska, A. Klin, and F. Volkmar. Automatic attention cueing through eye movement in 2-year-old children with autism. *Child Development*, 74:1108–1122, 2003.

M. G. Cline. The perception of where a person is looking. *American Journal of Psychology*, 80:41–50, 1967.

D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'00)*, Hilton Head Island, SC, June 2000.

V. Corkum and C. Moore. Development of joint visual attention in infants. In C. Moore and P. J. Dunham, editors, *Joint Attention: Its Origins and Role in Development*, pages 61–83. Erlbaum, Hillsdale, NJ, 1995.

V. Corkum and C. Moore. The origins of joint visual attention in infants. *Developmental Psychology*, 34:28–38, 1998.

L. Cosmides, J. Tooby, and H. Barkow. Evolutionary psychology and conceptual integration. In J. Barkow, L. Cosmides, and J. Tooby, editors, *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford University Press, New York, 1992.

S. Daly, K. Matthews, and J. Ribas-Corbera. Visual eccentricity models in face-based video compression. In *Proceedings of the IS&T/SPIE Conference on Human Vision and Electronic Imaging*, San Jose, CA, January 1999.

J. L. Dannemiller and B. K. Stevens. A critical test of infant pattern preference models. *Child Development*, 59:210–216, 1988.

M. Daprett, M. Davies, J. Pfeifer, A. Scott, M. Simgna, S. Bookheimer, and M. Iacoboni. Understanding emotions in others: Mirror neuron dysfunction in children with autism spectrum disorders. *Nature*, 1:28–30, 2006.

G. Dawson, A. N. Meltzoff, J. Osterling, J. Rinaldi, and E. Brown. Children with autism fail to orient to naturally occurring social stimuli. *Journal of Autism and Developmental Disorders*, 28:479–485, 1998.

G. Dawson, K. Toth, R. Abbott, J. Osterling, J. Munson, A. Estes, and J. Liaw. Early social attention impairments in autism: Social orienting, joint attention, and attention to distress. *Developmental Psychology*, 40:271–283, 2004.

P. Dayan and L. F. Abbott. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems.* MIT Press, Cambridge, MA, 2001.

P. Dayan, S. Kakade, and P. R. Montague. Learning and selective attention. *Nature Neuroscience Supplement*, 3:1218–1223, 2000.

G. O. Deák, R. Flom, and A. D. Pick. Perceptual and motivational factors affecting joint visual attention in 12- and 18-month-olds. *Developmental Psychology*, 36: 511–523, 2000.

G. O. Deák, Y. Wakabayashi, and H. Jasso. Attention sharing in human infants from 5 to 10 months of age in naturalistic conditions. In *Proceedings of the Third International Conference on Development and Learning (ICDL'04)*, La Jolla, CA, October 2004.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38, 1977.

D. C. Dennet. *The Intentional Stance.* MIT Press, Cambridge, MA, 1987.

A. Diamond. Differences between adult and infant cognition: Is the crucial variable presence or absence of language? In L. Weiskrantz, editor, *Thought Without Language*, pages 337–370. Oxford University Press, Oxford, 1988.

K. Doya. Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion in Neurobiology*, 10:732–739, 2000.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* Wiley-Interscience, New York, 2001.

P. J. Dunham and F. Dunham. Optimal social structures and adaptive infant behavior. In C. Moore and P. J. Dunham, editors, *Joint Attention: Its Origins and Role in Development*, pages 159–188. Erlbaum, Hillsdale, NJ, 1995.

J. L. Elman, E. A. Bates, M. H. Johnson, A. Karmiloff-Smith, D. Parisi, and K. Plunkett. *Rethinking Innateness: A Connectionist Perspective on Development.* MIT Press, Cambridge, MA, 1996.

T. Farroni, G. Csibra, F. Simion, and M. H. Johnson. Eye contact detection in humans from birth. *Proceedings of the National Academy of Sciences of the United States of America*, 99:9602–9605, 2002.

I. Fasel, G. O. Deák, J. Triesch, and J. Movellan. Combining embodied models and empirical research for understanding the development of shared attention. In *Proceedings of the International Conference on Development and Learning (ICDL'02)*, Cambridge, MA, June 2002.

S. Feinman. Social referencing in infancy. *Merrill-Palmer Quarterly*, 28:445–470, 1982.

J. H. Flavell. The development of inferences about others. In T. Mischel, editor, *Understanding Other Persons*, pages 66–116. Blackwell, Oxford, 1974.

J. H. Flavell. Development of children's knowledge about the mental world. *International Journal of Behavioral Development*, 24:15–23, 2000.

D. Frye. The origins of intention in infancy. In D. Frye and C. Moore, editors, *Children's Theories of Mind.* Erlbaum, Hillsdale, NJ, 1991.

D. Frye and C. Moore. *Children's Theories of Mind: Mental States and Social Understanding*. Erlbaum, Hillsdale, NJ, 1991.

M. Fujita and H. Kitano. Development of an autonomous quadruped robot for robot entertainment. *Autonomous Robots*, 5:7–18, 1998.

E. J. Gibson and N. Rader. Attention: The perceiver as performer. In G. Hale and M. Lewis, editors, *Attention and Cognitive Development*, pages 1–21. Plenum, New York, 1979.

J. J. Gibson and A. Pick. Perception of another person's looking behavior. *American Journal of Psychology*, 76:386–394, 1963.

A. Gopnik and A. N. Meltzoff. Imitation, cultural learning, and the origins of "theory of mind". *Behavioral and Brain Sciences*, 16:521–522, 1993.

L. Grover. *Comprehension of the Manual Pointing Gesture in Human Infants*. PhD thesis, University of Southamptom, 1988.

V. V. Hafner and F. Kaplan. Learning to interpret pointing gestures: Experiments with four-legged autonomous robots. In S. Wermter, G. Palm, and M. Elshaw, editors, *Biomimetic Neural Learning for Intelligent Robots, LNCS 3575*, pages 225–234. Springer Verlag, Heidelberg, Germany, 2005.

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

M. Hayhoe, M. Land, and A. Shrivastava. Coordination of eye and hand movements in a normal environment. *Investigative Ophthalmology and Visual Science*, 40: S380, 1999.

C. Heyes. Causes and consequences of imitation. *Trends in Cognitive Sciences*, 5: 253–261, 2001.

M. W. Hoffman, D. B. Grimes, A. P. Shon, and R. P. N. Rao. A probabilistic model of gaze imitation and shared attention. *Neural Networks*, 19:299–310, 2006.

B. M. Hood. Shifts of visual attention in the human infant: A neuroscientific approach. In L. Lipsitt and C. Rovee-Collier, editors, *Advances in Infancy Research, Vol. 9*, pages 163–216. Ablex Publishing, Norwood, NJ, 1995.

J. C. Houk, J. L. Davis, and D. G. Beiser, editors. *Models of Information Processing in the Basal Ganglia*. MIT Press, Cambridge, MA, 1995.

N. K. Humphrey. The social function of the intellect. In P. P. G. Bateson and R. A. Hinde, editors, *Growing Points in Ethology*. Cambridge University Press, Cambridge, 1976.

C. Hutt and C. Ounsted. The biological significance of gaze aversion with particular reference to the syndrome of infantile autism. *Behavioral Science*, 11:346–356, 1966.

M. Iacobini, I. Molnar-Szakacs, V. Gallese, G. Buccino, J. Mazziotta, and G. Rizzolatti. Grasping the intentions of others with one's own mirror neuron system. *Public Library of Science Biology*, 3:539–535, 2005.

M. Iacobini, R. Woods, M. Brass, H. Bekkering, J. Mazziotta, and G. Rizzolatti. Cortical mechanisms of human imitation. *Science*, 286:2526–2528, 1999.

L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40:1489–1506, 2000.

H. Jasso and J. Triesch. A virtual reality platform for modeling cognitive development. In *Proceedings of the Third International Conference on Development and Learning (ICDL'04)*, La Jolla, CA, October 2004.

H. Jasso, J. Triesch, and G. O. Deák. Using eye direction cues for gaze following - a developmental model. In *Proceedings of the Fifth International Conference on Development and Learning (ICDL'06)*, Bloomington, IN, May 2006a.

H. Jasso, J. Triesch, C. Teuscher, and G. O. Deák. A reinforcement learning model explains the development of gaze following. In *Proceedings of the Seventh International Conference on Cognitive Modeling (ICCM 2006)*, Trieste, Italy, April 2006b.

R. Jenkins, J. D. Beaver, and A. J. Calder. I thought you were looking at me! direction-specific aftereffects in gaze perception. *Psychological Science*, 17:506–514, 2006.

M. H. Johnson, M. I. Posner, and M. K. Rothbart. Facilitation of saccades toward a covertly attended location in early infancy. *Psychological Science*, 5:90–93, 1994.

A. Johnston. Spatial scaling of central and peripheral contrast sensitivity functions. *Journal of the Optical Society of America A*, 4:1583–1593, 1987.

A. Jolly. Lemur social behavior and primate intelligence. *Science*, 29:501–506, 1966.

S. S. Jones. Infants learn to imitate by being imitated. In *Proceedings of the Fifth International Conference on Development and Learning (ICDL'06)*, Bloomington, IN, May 2006.

M. I. Jordan. *Learning in Graphical Models.* MIT Press, Cambridge, MA, 1998.

R. E. Kalman. A new approach to linear filtering and prediction problem. *Journal of Basic Engineering*, 82:35–45, 1960.

K. Kaye. *The Mental and Social Lives of Babies: How Parents Create Persons*. University of Chicago Press, Chicago, 1982.

C. Keysers and D. Perrett. Demystifying social cognition: A hebbian perspective. *Trends in Cognitive Sciences*, 8:501–507, 2004.

H. Kim, G. York, G. Burton, E. Murphy-Chutorian, and J. Triesch. Design of an anthropomorphic robot head for studying autonomous development and learning. In *Proceedings of the 2004 IEEE International Conference on Robotics & Automation*, New Orleans, April 2004.

A. Klin, W. Jones, R. Schultz, and F. Volkmar. The enactive mind, or from actions to cognition: lessons from autism. *Philosophical Transactions of the Royal Society London Biological Science*, 358:345–360, 2003.

H. Kozima. Infanoid: A babybot that explores the social environment. In K. Dautenhahn, A. H. Bond, L. Canamero, and B. Edmonds, editors, *Socially Intelligent Agents: Creating Relationships with Computers and Robots*, pages 157–164. Kluwer Academic Publishers, Amsterdam, 2002.

H. Kozima and H. Yano. A robot that learns to communicate with human caregivers. In *Proceedings of the International Workshop on Epigenetic Robotics (EpiRob-2001)*, Lund, Sweden, September 2001.

B. Lau and J. Triesch. Learning gaze following in space: a computational model. In *Proceedings of the Third International Conference on Development and Learning (ICDL'04)*, La Jolla, CA, October 2004.

T. W. Lee, T. Wachtler, and T. J. Sejnowski. Color opponency is an efficient representation of spectral properties in natural scenes. *Vision Research*, 42: 2095–2103, 2002.

S. R Leekam, E. Hunnisett, and C. Moore. Targets and cues: Gaze following in children with autism. *Journal of Child Psychology and Psychiatry*, 39:951–962, 1998.

J. D. Lempers. Young children's production and comprehension of nonverbal deictic behaviors. *Journal of Genetic Psychology*, 135:93–102, 1979.

A. M. Leslie. ToMM, ToBy, and Agency: Core architecture and domain specificity. In L. Hirschfeld and S. Gelman, editors, *Mapping the Mind: Domain Specificity in Cognition and Culture*, pages 119–148. Cambridge University Press, New York, 1994.

K. Loveland and S. Landry. Joint attention and language in autism and developmental language delay. *Journal of Autism and Developmental Disorders*, 16: 335–349, 1986.

S. Maestro, F. Muratori, M. C. Cavallaro, F. Pei, D. Stern, B. Golse, and F. Palacio-Espasa. Attentional skills during the first 6 months of age in autism spectrum disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, 41:1239–1245, 2002.

A. N. Meltzoff. Imitation and other minds: The "like me" hypothesis. In S. Hurley and N. Chater, editors, *Perspectives on Imitation: From Neuroscience to Social Science*, pages 55–77. MIT Press, Cambridge, MA, 2005.

A. N. Meltzoff. 'like me': A foundation for social cognition. *Child Development*, 10:126–134, 2007.

A. N. Meltzoff and R. Brooks. Developmental changes in social cognition with an eye towards gaze following. In M. Carpenter and M. Tomasello (Chairs), editors, *Action-based measures of infants 'understanding of others' intentions and attention. Symposium conducted at the Bennial meeting of the International Conference on Infant Studies.* Chicago, 2004.

A. N. Meltzoff and R. Brooks. Eyes wide shut: The importance of eyes in infant gaze following and understanding other minds. In R. Flom, K. Lee, and D. Muir, editors, *Gaze Following: Its Development and Significance.* Erlbaum, Mahwah, NJ, 2006.

A. N. Meltzoff and A. Gopnik. The role of imitation in understanding persons and developing a theory of mind. In H. Tager-Flusberg S. Baron-Cohen and D. Cohen, editors, *Understanding Other Minds: Perspectives from Autism*, pages 335–366. Oxford University Press, Oxford, 1993.

A. N. Meltzoff and M. K. Moore. Explaining facial imitation: A theoretical model. *Early Development and Parenting*, 6:179–192, 1977.

G. Metta, G. Sandini, L. Natale, L. Craighero, and L. Fadiga. Understanding mirror neurons: a bio-robotic approach. *Interaction Studies*, 7:197–232, 2006.

H. Moll and M. Tomasello. Level 1 perspective-taking at 24 months of age. *British Journal of Developmental Psychology*, 24:603–616, 2006.

C. Moore. Gaze following and the control of attention. In P. Rochat, editor, *Early Social Cognition.* Erlbaum, Hillsdale, NJ, 1999.

C. Moore. *The Development of Commonsense Psychology.* Erlbaum, Mahwah, NJ, 2006.

C. Moore and V. Corkum. Social understanding at the end of the first year of life. *Developmental Review*, 14:349–372, 1994.

C. Moore and P. J. Dunham. *Joint Attention: Its Origins and Role in Development.* Erlbaum, Hillsdale, NJ, 1995.

P. Morissette, M. Ricard, and T. G. Dcarie. Joint visual attention and pointing in infancy: A longitudinal study of comprehension. *British Journal of Developmental Psychology*, 13:163–175, 1995.

P. Mundy, M. Sigman, and C. Kasari. A longitudinal study of joint attention and language development in autistic children. *Journal of Autism and Developmental Disorders*, 20:115–128, 1990.

L. Murray and C. Trevarthen. Emotion regulation of the interactions between two-month-olds and their mothers. In T. M. Field and N. A. Fox, editors, *Social Perception in Infants*, pages 89–111. Ablex, Norwood, NJ, 1985.

Y. Nagai. Joint attention development in infant-like robot based on head movement imitation. In *Proceedings of the Third International Symposium on Imitation in Animals and Artifacts (AISB'05)*, Hatfield, UK, April 2005a.

Y. Nagai. Self-other motion equivalence learning for head movement imitation. In *Proceedings of the Fourth International Conference on Development and Learning (ICDL'05)*, Osaka, Japan, July 2005b.

Y. Nagai, M. Asada, and K. Hosoda. Learning for joint attention helped by functional development. *Advanced Robotics*, 20:1165–1181, September 2006.

Y. Nagai, K. Hosoda, A. Morita, and M. Asada. A constructive model for the development of joint attention. *Connection Science*, 20:211–229, 2003.

E. Oztop and M. Arbib. Schema design and implementation of the grasp-related mirror neuron system. *Biological Cybernetics*, 87:116–140, 2002.

E. Oztop, M. Kawato, and M. Arbib. Mirror neurons and imitation: A computationally guided review. *Neural Networks*, 19:254–271, 2006.

K. A. Pelphrey, J. P. Morris, and G. McCarthy. Neural basis of eye gaze processing deficits in autism. *Brain*, 128:1038–1048, 2005.

J. Perner. *Understanding the Representational Mind.* MIT Press, Cambridge, MA, 1991.

D. Perret, M. Harries, A. Mistlin, J. Hietanen, P. Benson, R. Bevan, S. Thomas, M. Oram, J. Ortega, and K. Bierley. Social signals analyzed at the single cell level: Someone is looking at me, something touched me, something moved! *International Journal of Comparative Psychology*, 4:25–55, 1990.

D. Perret, P. Smith, D. Potter, A. Mistlin, A. Head, A. Milner, and M. Jeeves. Visual cells in the temporal cortex. *Experimental Brain Research*, 47:329–342, 1985.

J. Piaget and B. Inhelder. *The Child's Conception of Space (translated by F. J. Langdon and J. L. Lunzer).* Routledge and Kegan Paul, London, 1948/1956.

G. Pobric and A. D. C. Hamilton. Action understanding requires the left inferior frontal cortex. *Current Biology*, 16:524–529, 2006.

D. Premack and G. Woodruff. Does the chimpanzee have 'a theory of mind'? *Behavioral and Brain Sciences*, 4:515–526, 1978.

V. Reddy. Playing with others expectations: Teasing and mucking about in the first year. In A. Whiten, editor, *Natural Theories of Mind: Evolution, Developmnet, and Simulation of Everyday Mindreading*. Blackwell, Oxford, 1991.

B. Repacholi and A. Gopnik. Early understanding of desires: Evidence from 14 and 18-month-olds. *Developmental Psychology*, 33:12–21, 1997.

G. Rizzolatti. The mirror neuron system and its function in humans. *Anatomy and Embryology*, 210:419–421, 2005.

G. Rizzolatti and L. Craighero. The mirror-neuron system. *Annual Review of Neuroscience*, 27:169–192, 2004.

S. Roberts and H. Pashler. How persuasive is a good fit? A comment on theory testing. *Psychological Review*, 107:358–367, 2000.

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

B. Scassellati. Theory of mind for a humanoid robot. *Autonomous Robots*, 12: 13–24, 2002.

W. Schultz, P. Dayan, and P. R. Montague. A neural substrate of prediction and reward. *Nature*, 275:1593–1599, 1997.

J. C. Stanley. Computer simulation of a model of habituation. *Nature*, 261:146–148, 1976.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

D. Tantam, D. Holmes, and C. Cordess. Nonverbal expression in autism of asperger type. *Journal of Autism and Developmental Disorders*, 23:111–133, 1993.

D. Terzopoulos, X. Tu, and T. Grzeszczuk. Artificial fishes: Autonomous locomotion, perception, behavior and learning in a simulated physical world. *Artificial Life*, 1:327–351, 1994.

C. Teuscher and J. Triesch. To care or not to care: analyzing the caregiver in a computational gaze following framework. In *Proceedings of the Third International Conference for Development and Learning (ICDL'04)*, La Jolla, CA, October 2004.

C. Teuscher and J. Triesch. To each his own: The caregiver's role in a computational model of gaze following. *Neurocomputing*, in press.

E. Thelen and L. B. Smith. *A Dynamic Systems Approach to the Development of Cognition and Action.* MIT Press, Cambridge, MA, 1994.

A. L. Thomaz, M. Berlin, and C. Breazeal. An embodied computational model of social referencing. In *Proceedings of Fourteenth IEEE Workshop on Robot and Human Interactive Communication (Ro-Man05)*, Nashville, TN, August 2005.

M. Tomasello. Joint attention as social cognition. In C. Moore and P. J. Dunham, editors, *Joint Attention: Its Origins and Role in Development*, pages 103–130. Erlbaum, Hillsdale, NJ, 1995.

M. Tomasello, J. Call, and B. Hare. Chimpanzees understand psychological states - the question is which ones and to what extent. *Trends in Cognitive Sciences*, 4:153–156, 2003.

M. Tomasello, A. C. Kruger, and H. H. Ratner. Cultural learning. *Behavioral and Brain Sciences*, 16:495–511, 1993a.

M. Tomasello and H. Rakoczy. What makes human cognition unique? From individual to shared to collective intentionality. *Mind and Language*, 18:121–147, 2003.

M. Tomasello, E. S. Savage-Rumbaugh, and A. C. Kruger. Imitative learning of actions on objects by children, chimpanzees, and enculturated chimpanzees. *Child Development*, 64:1688–1705, 1993b.

C. Trevarthen. Instincts for human understanding and for cultural cooperation: Their development in infancy. In M. von Cranach, K. Foppa, W. Lepenies, and D. Ploog, editors, *Human Ethology: Claims and Limits of a New Discipline*, pages 530–571. Cambridge University Press, Cambridge, 1979.

C. Trevarthen. Predispositions to cultural learning in young infants. *Behavioral and Brain Sciences*, 16:534–535, 1993.

C. Trevarthen and P. Hubley. Secondary intersubjectivity: Confidence, confiding and acts of meaning in the first year. In A. Lock, editor, *Action, Gesture, and Symbol*, pages 183–229. Academic Press, London, 1978.

J. Triesch, H. Jasso, and G. O. Deák. Emergence of mirror neurons in a model of gaze following. In *Proceedings of the Fifth International Conference on Development and Learning (ICDL'06)*, Bloomington, IN, May 2006a.

J. Triesch, H. Jasso, and G. O. Deák. Emergence of mirror neurons in a model of gaze following. *Adaptive Behavior*, in press.

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

V. Virsu and J. Rovamo. Visual resolution, contrast sensitivity, and the cortical magnification factor. *Experimental Brain Research V*, 37:475–494, 1979.

C. von Hofsten, E. Dahlström, and Y. Fredriksson. 12-month-old infants' perception of attention direction in static video images. *Infancy*, 8:217–231, 2005.

J. S. Watson. Smiling, cooing, and the game. *Merrill-Palmer Quaterly*, 18:323–339, 1972.

C. Weber, S. Wermter, and M. Elshaw. A hybrid generative and predictive model of the motor cortex. *Neural Networks*, 19:339–353, 2001.

H. M. Wellman. Early understanding of mind: The normal case. In S. Baron-Cohen, H. Tager-Flusberg, and D. J. Cohen, editors, *Understanding Other Minds: Perspectives from Autism*, pages 67–105. Oxford University Press, Oxford, 1991.

J. Weng, J. McClelland, A. Pentland, O. Sporns, I. Sotckman, M. Sur, and E. Thelen. Autonomous mental development by robots and animals. *Science*, 291: 599–600, 2001.

H. Werner and B. Kaplan. *Symbol Formation*. Wiley, New York, 1963.

A. Whiten. *Natural Theories of Mind. Evolution, Development, and Simulation of Everyday Mindreading*. Blackwell, Oxford, 1991.

A. Whiten and J. Perner. *Fundamental Issues In the Multidisciplinary Study of Mindreading*. Basil Blackwell, Oxford, 1991.

Y. Wu, K. Toyama, and T. Huang. Wide-range, person - and illumination-insensitive head orientation estimation. In *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000*, Grenoble, France, March 2000.

C. Yu, H. Zhang, and L. B. Smith. Learning through multimodal interaction. In *Proceedings of the Fifth International Conference on Development and Learning (ICDL'06)*, Bloomington, IN, May 2006.

H. Zhang, C. Yu, and L. B. Smith. An interactive virtual reality platform for studying embodied social interaction. In *Proceedings of the CogSci06 Symposium Toward Social Mechanisms of Android Science*, Vancouver, Canada, July 2006.