Emergence of Mirror Neurons in a Model of Gaze Following

short title: Mirror Neurons in a Model of Gaze Following

Jochen Triesch

Frankfurt Institute for Advanced Studies, Johann Wolfgang Goethe University

Max-von-Laue-Str. 1, 60438 Frankfurt am Main, Germany

triesch@fias.uni-frankfurt.de

phone: +49 69 798-47531

fax: +49 69 798-47611

and

Department of Cognitive Science, UC San Diego

9500 Gilman Drive, La Jolla, CA 92093-0515, USA

Hector Jasso

Department of Computer Science and Engineering, UC San Diego

9500 Gilman Drive, La Jolla, CA 92093, USA

hjasso@ucsd.edu

Gedeon O. Deák

Department of Cognitive Science, UC San Diego

9500 Gilman Drive, La Jolla, CA 92093-0515, USA

deak@cogsci.ucsd.edu

Abstract

Gaze following is the ability to re-direct one's gaze to the location where another agent is looking. We present a computational model of how human infants or other agents may acquire gaze following by learning to predict the locations of interesting sights from the looking behavior of other agents through reinforcement learning. The model accounts for many findings about the development of gaze following in human infants. During learning, the model develops pre-motor representations that exhibit many properties characteristic of mirror neurons, but they are specific to looking behaviors. The existence of such a new class of mirror neurons is the main prediction of our model. The model also offers a parsimonious account of how these and possibly other mirror neurons may acquire their special response properties. In this account, visual representations of other agents' actions become associated with pre-motor neurons that represent the intention to perform corresponding actions. The model also demonstrates how this development may be obstructed in autism spectrum disorder, giving rise to specific physiological and anatomical differences in the mirror system.

Keywords: gaze following, shared attention, mirror neuron, imitation, autism, reinforcement learning

1 Introduction

1.1 What is Gaze Following?

Gaze following is the ability to look where somebody else is looking. This skill is considered a foundational component of humans' social interaction abilities and belongs to the family of attention sharing behaviors. Gaze following is also present in a number of other species and is of great importance for social robots that must interact with people. In humans, gaze following emerges in a progressive fashion during the first two years of life (Scaife & Bruner, 1975; Butterworth & Jarrett, 1991). While pre-cursors of gaze following can be observed very early (D'Entremont, Hains, & Muir, 1997; Farroni, Massaccesi, Pividori, & Johnson, 2004), some gaze following behaviors do not emerge until 18 months or later. For example, whereas 9-month-olds will follow an adult's gaze only to targets inside their visual field (Flom, Deák, Phill, & Pick, 2003), 12-month-olds will follow gaze to targets in their periphery or behind them (Deák, Flom, & Pick, 2000) although this skill is consolidated between 12 and 18 months (Butterworth & Jarrett, 1991; Deák et al., 2000). Similarly, while young infants will easily be "fooled" by additional distractor objects which are not being looked at, older infants are more accurate in estimating the correct target of others' gaze (Butterworth & Jarrett, 1991). Finally, there is an interesting development in the kinds of visual cues that infants use for gaze following. While young infants seem to mostly follow others' head movements, they will later become more sensitive to the status of the eyes of the other person (Corkum & Moore, 1995; Caron, Butler, & Brooks, 2002). There has been much interest in gaze following in recent years, because researchers believe that infants' developing gaze following capacities are precursors of their develop-

ing understanding of others as perceiving, intelligent agents. In fact, many gaze following experiments have been designed specifically with the goal of elucidating the nature of the infant's understanding of other people.

1.2 Why does Gaze Following Emerge?

Maybe the most fundamental question about gaze following is why it emerges at all. Initial accounts had a nativist flavor, explaining the phenomenon in terms of specific innate *modules* that mature during development (Leslie, 1987; Baron-Cohen, 1995). Modern accounts have emphasized the role of learning processes as the infant interacts with its social environment (Moore & Corkum, 1994; Corkum & Moore, 1995; Fasel, Deák, Triesch, & Movellan, 2002). Over the last years, several groups have been working on computational models that try to test the plausibility of a learning account of gaze following (Matsuda & Omori, 2001; Nagai, Hosoda, Morita, & Asada, 2003; Carlson & Triesch, 2004; Lau & Triesch, 2004; Hoffman, Grimes, Shon, & Rao, 2006; Nagai, 2005; Triesch, Teuscher, Deák, & Carlson, 2006; Teuscher & Triesch, 2006). These models attempt to account for various aspects of the infant's developing gaze following abilities and explain this development in neural terms. It is beneficial to keep such models as simple as possible. In doing so, one does not deny that infants will ultimately develop very sophisticated representations of themselves and others. Rather, one would like to clarify *how* these representations may emerge, how they may be built on top of earlier and more primitive representations, and what the underlying developmental processes may be. For this it is best to start with relatively simple models and to extend and refine them as needed. While good progress has been made along these lines, there is presently no model that fully captures all aspects of

the development of gaze following abilities during the first 18 month of life.

1.3 Gaze Following as Imitation

Gaze following can be linked to imitation. In its most general sense, imitation occurs when an individual observes another's behavior and replicates it (Tomasello, 1999). In this sense, gaze following can be viewed as imitation (Nagai, 2005; Hoffman et al., 2006). The behavior that is being observed is another's gaze shift to a particular location in space, and this behavior is replicated. Most authors use more strict definitions for "true" imitation, however, and might prefer to consider acts of gaze following as a form of *response facilitation* because the copied behavior (a gaze shift to a certain location) is not novel but already part of the agent's behavioral repertoire (e.g. Schaal, 1999). In addition, the motor patterns of the model will usually be quite different from that of the person following gaze. For example, when both are facing each other then the correct response to the model turning to her right side will be the imitator turning to her own left side. So rather than copying a specific pattern of motor activation, i.e. mimicking a specific head turn, gaze following implies the emulation of the result or goal of the model's action, which is to fixate a certain object at a particular position in space. At any rate, this link between gaze following and imitation implies that gaze following may emerge in a similar way and for similar reasons as the emergence of other imitative behaviors. It has even been suggested that gaze following may be a necessary condition for certain forms of imitation to emerge (Kumashiro et al., 2003). The link between gaze following and imitation also implies that the neural basis of gaze following may be closely related to the neural basis of other imitative behaviors.

1.4 Imitation and the Mirror Neuron System

Mirror neurons are a class of pre-motor neurons originally found in macaque area F5. There is converging evidence from fMRI, EEG, and MEG studies for the existence of a similar system of mirror neurons in humans, where direct observation of individual mirror neurons is impossible with today's experimental techniques (Rizzolatti & Craighero, 2004). Their defining characteristic is that they become activated when the animal performs an action such as reaching for an object and grasping it or when the animal sees another agent perform the same or a similar action. Because of this property, it has been suggested that mirror neurons may play a role in a number of social-cognitive functions (Rizzolatti, 2005) including understanding others' actions (Pobric & Hamilton, 2006), imitation (e.g. Iacoboni et al., 1999), intention understanding (Iacoboni et al., 2005), and empathy (review in Gallese, Keysers, & Rizzolatti, 2004). Some mirror neurons can be triggered through different modalities. For example, a mirror neuron in monkey pre-motor cortex may respond to seeing the action of tearing paper or to hearing the same action being performed (Kohler et al., 2002). Interestingly, some mirror neurons in the ventral pre-motor cortex of macaques will respond to an action that can be inferred, but is not fully visible (Umiltà et al., 2001).

It is easy to imagine the role that mirror neurons may play in imitation or, more precisely, response facilitation, although direct evidence is still lacking. If another agent is observed performing an action, then this leads to the activation of a population of mirror neurons that code for this action. Because of the motor properties of mirror neurons, this representation of the other agent's action might be used to plan execution of a corresponding action. One important question in this context is what "corresponding" means, i.e. at what

"level" the action is represented. Current evidence is most consistent with the interpretation that it is not the detailed motor pattern that is encoded by mirror neurons but the general type, purpose, and functional completion of the action.

What is also unclear is how mirror neurons arrive at their specific response properties. We find it very unlikely that a sophisticated mirror system could be innate, in the sense of a detailed pre-specified connection pattern for every neuron. Rather, learning processes must play an important role both in the formation of the mirror system and in the development of imitation (Heyes, 2001; Keysers & Perrett, 2004; Brass & Heyes, 2005; Jones, 2006). At present, however, there have been no studies investigating whether mirror neurons are present in infants of different ages. Consequently, nothing is known about what experiences and interactions with the environment are necessary or sufficient for the emergence of mirror neurons.

1.5 Contributions of this Work

In the following we present a new model of the emergence of gaze following that can account for a wide range of experimental findings and is based on neurally plausible reinforcement learning mechanisms. Preliminary versions of this model have appeared in (Jasso, Triesch, Teuscher, & Deák, 2006; Triesch, Jasso, & Deák, 2006). In our model, the developing infant learns to associate visual representations of other agents' gaze shifts with pre-motor representations for corresponding gaze shifts. As a consequence, the model develops internal pre-motor representations that share important properties with mirror neurons. Thus, the model predicts the existence of a new class of mirror neurons for looking behaviors that has not been observed experimentally. In addition, it offers a new account of

how neurons with mirror properties can be learned "from scratch" that is distinct from and complementary to previous proposals of how mirror neurons could emerge through learning. Finally, we simulate alterations of this development in autism spectrum disorders giving rise to physiological and anatomical differences in the predicted population of mirror neurons.

2 Model Description

Our model is an extension of our earlier modeling work (Carlson & Triesch, 2004; Lau & Triesch, 2004; Triesch, Teuscher, et al., 2006). A major novelty in the new model is that it allows modeling of several spatial aspects of the gaze following problem. In the model, an infant and a caregiver interact with a number of more or less visually salient objects as illustrated in Fig. 1. The caregiver and infant are looking back and forth between the objects and each other, driven by visual saliency and habituation. During the process, the infant learns to predict the locations of salient objects based on the looking behavior of the caregiver. Time is running in discrete steps corresponding to roughly half a second. All parameters of the model and their allowed ranges and default values are summarized in Table 1.

Figure 1. `<< about here >>`

2.1 Infant, Caregiver, and Environment

The interaction of infant and caregiver in their environment is organized as follows. There are periods when the caregiver is present alternating with periods when the infant is alone with the objects (the caregiver has left the room). The duration of these periods are drawn from geometric distributions with means $\bar{T}_{\text{present}} = 120$ time steps and $\bar{T}_{\text{absent}} = 120$

Table 1: << about here >>

time steps, respectively. When the caregiver is present, the infant and caregiver are in fixed locations facing each other with a 40 cm separation between them.

At any time a random number of objects, drawn from a geometric distribution with mean $\bar{\Phi}_o = 4$, will be present. Their locations are drawn from a 2-dim. Gaussian distribution centered at the infant with a standard deviation $\sigma_o = 0.5$m. The objects will remain stationary for a random number of time steps drawn from a geometric distribution with mean $\bar{T}_{\text{objects}} = 10$. After that they are replaced by a new set of objects.

Associated with caregiver, infant, and objects are visual saliencies Φ_C, Φ_I, and Φ_{o_j}, respectively. The saliencies of the objects Φ_{o_j} are drawn from an exponential probability distribution with mean $\bar{\Phi}_o = 1$. We set the saliency of the caregiver and the infant to $\Phi_C = \Phi_I = 2$. This implies that most objects have a lower saliency than that of the infant and caregiver but higher saliencies can also occur.

2.2 Infant Visual System

The infant's visual input is processed by three different sub-systems (see Fig. 2, left) that serve different functions: a saliency map that represents where visually interesting stimuli are located, a representation of the head pose of the caregiver, and a subsystem for representing the gaze direction of the caregiver. The infant can only directly perceive objects that fall inside its field of view, which extends to $\pm 90^\circ$ around the infant's current viewing direction. However, the infant may have a memory trace for objects outside its current field of view. In the following, we refer to locations that fall inside the infant's current field of view as *visible* and others as *not visible*.

Figure 2. `<< about here >>`

The first component of the infant's visual system is a *saliency map* that allows the infant to represent the locations of interesting visual targets. Such representations are commonly believed to be involved in the planning of eye movements in the primate visual system. In our model, the *saliency map* $\mathbf{s} = (s_1, \ldots, s_{64})^T$ indicates the presence of visual saliency in a body-centered coordinate system around the infant. It is discretized into 64 different regions in space, corresponding to 16 uniformly spaced heading ranges and 4 depth ranges covering depths of up to 0.3m, 0.6m, 0.9m, and beyond, respectively. Our assumption of a body-centered representation (in contrast to a retinotopic one) may not be physiologically accurate but it frees us from having to model coordinate transformations between different coordinate systems. It is an interesting question in its own right when and how infants learn to compute certain coordinate transformations, but this question is beyond the scope of this paper. The total activation s_i at location i in the saliency map is calculated as:

$$s_i(t) = \begin{cases} \sum_j f_j(t)\phi_{o_j}(t) & : \text{location } i \text{ visible} \\ d s_i(t-1) & : \text{location } i \text{ not visible} \end{cases}, \quad (1)$$

where $d = 0.5$ is a factor that determines the speed of decay of the memory trace of the saliency at location i. The sum runs only over the objects j present in location i possibly including the caregiver. $\phi_{o_j}(t)$ is the *habituated saliency* of object j (explained below), and $f_j(t)$ is a *foveation factor*. This factor reduces the saliency of a visible object j the farther the object is from the infant's line of sight, i.e. the more peripherally it is located within the infant's current field of view. Specifically, we define $f_j(t) = \exp(-{\theta_{o_j}}^2/\sigma_F^2)$, where θ_{o_j} is the angle between the infant's line of sight and the object and $\sigma_F = 180°$ determines the

range of the attenuation.

Habituation further decreases the perceived saliency of an object. The *habituated saliency* $\phi_{o_j}(t)$ of object j is an attenuated version of the original saliency. The infant habituates separately to each object according to the discretized version of a habituation model proposed by Stanley (Stanley, 1976):

$$\tau_H \frac{d\phi_{o_j}(t)}{dt} = \alpha_H(\Phi_{o_j} - \phi_{o_j}(t)) - S_{o_j}(t) \ , \tag{2}$$

where $\phi_{o_j}(t)$ is object j's habituated saliency at time t and Φ_{o_j} is 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 $\alpha_H \geq 1$ controls down to what level the saliency can be reduced. More precisely, if the infant kept looking at an object forever, its saliency would eventually be reduced to a fraction $(\alpha_H - 1)/\alpha_H$. An equivalent formula applies for $\phi_C(t)$, the habituated saliency of the caregiver.

Next to the saliency map, the infant visual system extracts information about the head pose and gaze direction of the caregiver. 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, & 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 (Jasso & Triesch, 2006).

The second component of the infant visual system estimates the *caregiver head direction*. It is represented by a vector $\mathbf{h} = (h_1, \ldots, h_{16})^T$ that indicates 16 possible caregiver head directions as perceived by the infant. Heading ranges are similar to those in $\mathbf{s}$. If the

infant is looking at the caregiver, the h_i corresponding to the caregiver's head direction is set to 1. Human infants' sensitivity to different head poses is currently unknown, but it is clear that some rudimentary sensitivity is already present in 1-month-olds (Sai & Bushnell, 1989). If the infant is not looking at the caregiver, then the values of $\mathbf{h}$ are calculated by multiplying the previous value by the memory constant d, as in the calculation of $\mathbf{s}$. Concretely, h_i is given by:

$$h_i(t) = \begin{cases} 1 & : \text{caregiver visible and looking in direction } i \\ 0 & : \text{caregiver visible and not looking in direction } i \\ dh_i(t-1) & : \text{caregiver not visible} \end{cases} . \quad (3)$$

The third component of the infant visual system estimates the *caregiver eye direction*. It is represented by a vector $\mathbf{e} = (e_1, \ldots, e_{16})^T$ which is similar to $\mathbf{h}$, but computed based on the caregiver's direction of gaze rather than the caregiver's head orientation. The only other difference is that when the caregiver is present and within the infant's field of view, but turning her back to the infant, all e_i are set to zero. This reflects the fact that when the caregiver is facing backwards with respect to the infant, her eyes are not visible.

We can summarize the complete state of the infant's visual system by a *state vector* $\mathbf{u}$ that is a concatenation of $\mathbf{s}$, $\mathbf{h}$, and $\mathbf{e}$: $\mathbf{u} = (\mathbf{s}^T, \mathbf{h}^T, \mathbf{e}^T)^T$. The dimensionality of $\mathbf{u}$ is the sum of the dimensions of $\mathbf{s}$, $\mathbf{h}$, and $\mathbf{e}$, which is $N_s = 64 + 16 + 16 = 96$. The infant has to learn how to map this sensory representation onto appropriate behaviors.

2.3 Reinforcement Learning Model

The infant model learns through a biologically plausible reinforcement learning scheme (Sutton & Barto, 1998). In particular, the infant model is formulated as a so-called actor-

critic architecture. This is a popular approach for modeling skill acquisition in agents that has two separate structures — the actor and the critic. The critic learns to evaluate how "good" it is to be in a certain situation, taking into account any likely future actions and their consequences. The actor decides how to behave in a certain situation, i.e. it maps a representation of the current state of the world onto probabilities for selecting the next action.

The learning process tries to optimize the infant's policy, i.e. the way the actor maps sensory states onto different gaze shifts in order to maximize the long-term reward obtained by the infant. Learning in both the actor and the critic is driven by a so-called temporal difference (TD) error signal, that is calculated by the critic. The TD error has been associated with the activity of dopaminergic neurons in the midbrain (W. Schultz, Dayan, & Montague, 1997).

In our model, the state vector $\mathbf{u}$ of the infant's visual system serves as input to the actor-critic reinforcement learning system. The only possible actions the actor can produce are gaze shifts to $N_a = 64$ locations represented in a body-centered coordinate system congruent to that of the saliency map $\mathbf{s}$.

Reward is obtained as the saliency of the position where attention is directed to after a gaze shift has been made and $\mathbf{s}$ has been updated with the result of the action (value of $\mathbf{s}$ corresponding to the depth/heading of the selected gaze shift a). The definition of saliency as reward is based on studies of infant visual expectations and the organization of their behavior around these expectations (Haith, Hazan, & Goodman, 1988).

In order to maximize its reward, the infant utilizes a standard actor-critic learning scheme. The *critic* (see Fig. 2, upper right) approximates the value of the current state

as $v(t) = \mathbf{w}^T(t)\mathbf{u}(t)$, where $\mathbf{w}(t) = (w_1(t), w_2(t), \ldots, w_{N_s}(t))^T$ is a weight vector. During learning, the weight vector $\mathbf{w}(t)$ is updated according to:

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

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

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

where $r(t)$ is the reward after taking an action at time t, $v(t+1)$ is the estimated value of the new state after taking the action, and $0 \leq \gamma \leq 1$ is the reward discount factor.

The *actor* (see Fig. 2, lower right) specifies the action to be taken, directing the infant's attention to one of the N_a possible locations. It relies on a pre-motor representation $\mathbf{m}(t) = (m_1(t), m_2(t), \ldots, m_{N_a}(t))^T$ that learns to encode how desirable gaze shifts to each of the N_a locations are in any given situation. Formally, we define $\mathbf{m} = \mathbf{M}\mathbf{u}$, where $\mathbf{M}$ is an N_a-by-N_s weight matrix. Based on the pattern of pre-motor activations, an action $a(t)$ is chosen probabilistically such that a higher value of m_a increases the chances of selecting action a. This happens according to a *softmax* decision rule:

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

where m_a is the action value corresponding to action a and β is an "inverse temperature" parameter which increases exploitation versus exploration with a larger value. $\mathbf{M}$ is updated according to (Dayan & Abbott, 2001):

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

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

2.4 Caregiver Behavior

The behavior of the caregiver is very important for the success of the infant's learning process. At each time step, the caregiver looks at the most salient object, where saliency is mediated by the same habituation mechanism (with identical parameters) as in the infant's visual system. The caregiver's head direction can be slightly offset from that of the eyes according to a Gaussian distribution with mean zero and standard deviation $\sigma_C = 5°$. This offset is recalculated for every gaze shift of the caregiver. This reflects the fact that eyes and head are not always aligned because the eyes can move inside the head, and corresponds to values observed in naturalistic settings (Hayhoe, Land, & Shrivastava, 1999).

3 Experiments and Results

In the following, we describe two experiments. Experiment 1 demonstrates the normal emergence of gaze following in the model and shows how the pre-motor layer **m** acquires mirror properties. Experiment 2 shows how a disinterest in the face of the caregiver as seen in autism can slow down or even abolish this development, preventing the emergence of mirror properties in layer **m**.

Each experiment starts with all weights set to zero. The model is simulated for a total of 900,000 time steps, which corresponds to 125 hours of non-stop interaction if we associate 1 time step with half a second. The caregiver is present approximately half of the time. At regular intervals, we interrupt the learning process to evaluate the network's

behavior in a set of tests. Table 1 summarizes the values of all parameters used in the experiments (unless stated otherwise below).

3.1 Experiment 1: Normal Emergence of Gaze Following and Formation of Mirror Neurons

During the simulation, the weights $\mathbf{w}$ and $\mathbf{M}$ of the critic and the actor gradually change according to (4) and (7), as the infant model improves its behavior. The connections from layer $\mathbf{s}$ to layer $\mathbf{m}$ quickly develop a characteristic one-to-one mapping: the infant learns how to make accurate saccades to salient objects in its field of view. This does not require the presence of the caregiver but also happens in her absence. At the same time, the connections from $\mathbf{h}$ and $\mathbf{e}$ to the pre-motor area develop a characteristic pattern. The infant learns that specific head poses and eye directions of the caregiver predict rewarding stimuli in certain locations. Since the caregiver tends to look at the most salient object, these locations happen to lie along the caregiver's line of sight. Note, however, that the infant could also learn any other reliable association between the caregiver's cues and the locations of rewarding sights. Figure 3 shows all learned connection weights from layers $\mathbf{h}$ and $\mathbf{e}$ to layer $\mathbf{m}$ at the end of a simulation. Each pre-motor unit is strongly activated by just one or two units in layer $\mathbf{h}$ and layer $\mathbf{e}$. The only exception is pre-motor unit m_1 (bottom row in the figure), which corresponds to looking at the caregiver. Conversely, however, each unit in layer $\mathbf{h}$ and layer $\mathbf{e}$ typically activates several units in $\mathbf{m}$ that reflect the corresponding line of sight. This is illustrated more clearly in Fig. 4 for two example units.

Figure 3. `<< about here >>`

The model's behavior nicely reflects the developmental progression from rudimentary to more sophisticated gaze following observed in human infants (see Jasso et al., 2006 for

Figure 4. `<< about here >>`

a preliminary account). This development also results in an interesting transition in the infant's behavior. While the initial behavior is mostly driven by visual saliency (bottom-up attention), the later looking behavior becomes increasingly driven by top-down predictions about the locations of rewards.

At the end of the learning process, we continued the simulation for an additional 10^6 time steps with all connection weights being fixed while we "recorded" from units in the pre-motor layer. We found that the model neurons in layer **m** share many characteristics with classical mirror neurons. A unit in this layer will usually be active during the execution of a gaze shift to a certain location in space. This is because the probability of performing such a gaze shift is directly related to the activation of the unit, as described by (6). Figure 5 illustrates this for two example units. For both units we estimate the conditional probability of the units' activations given that a gaze shift to their locations is/is not made. Frequently, the units are highly activated when a gaze shift to their location occurs. When a gaze shift to a different location is made, high activity is observed only rarely.

The units in **m** also have interesting sensory properties. In particular, they will be active when the infant observes the caregiver looking in the corresponding direction. This is due to the learned connection weights from the representation of the caregiver's head and eyes in layers **e** and **h** to the pre-motor units in **m**. Figure 6 illustrates the sensory properties of the two pre-motor units A and B. We consider a situation where the infant is looking at the caregiver, while the caregiver either looks in the direction of the pre-motor units' locations, or in the direction of the corresponding location on the opposite side of

Figure 5. `<< about here >>`

the room. For unit A we also distinguish the cases that a salient object ($\Phi_o = 0.5$) is or is not present at the unit's location. For unit B, this distinction makes no difference, because unit B is outside of the infant's visual field in this situation (compare Fig. 3). Unit A is strongly activated by a salient object in its location, but this activation will be increased when the caregiver also looks in this direction. Even in the absence of any salient stimulus in its location, unit A will be activated if the caregiver looks in this direction, albeit less strongly. The result is similar for unit B, which becomes activated when the caregiver is looking in its direction.

The combination of being active during execution and observation of a motor act is the defining characteristic of mirror neurons. Clearly, the neurons in layer **m** can be viewed as mirror neurons. At the same time, these neurons are not merely motor neurons. The model will not always perform a gaze shift when the corresponding pre-motor neuron is activated. Instead, the pre-motor neurons in **m** only represent a plan or proposal to perform a certain gaze shift from which the action selection mechanism will choose one. This means that the activation of a pre-motor unit due to a salient stimulus or the gaze shift of another agent does not automatically lead to the corresponding gaze shift. Instead, multiple such action plans will usually compete. In addition, execution of any action may be inhibited by additional brain structures which we have not included in our model.

Note that area F5 mirror neurons selective for grasping typically do not respond to just the presence of a visual stimulus alone, such as a graspable object in the field of view without a grasping hand, even if that object is of interest to the animal. In this respect, the

Figure 6. `<< about here >>`

mirror neurons in our model behave differently. A salient visual object to which the model is not habituated will be sufficient to activate a neuron in layer m (if it is visible or only remembered). This activation will be stronger, however, if the model also sees the caregiver looking in the direction of the object.

3.2 Experiment 2: Modeling Differences in Autism Spectrum Disorder

Autism spectrum disorder is a pervasive developmental disorder that is characterized by abnormal communication ability, patterns of interests and behaviors, and social interaction behavior (Dawson et al., 2004) including imitation (Williams, Whiten, & Singh, 2004). Deficits in attention sharing are among the earliest behavioral predictors of the social and language deficits in autism (Osterling & Dawson, 1994). Interestingly, autistic individuals exhibit a disinterest in social stimuli, in particular faces (Adrien, Lenoir, Martineau, Perrot, & al., 1993; Chawarska, Klin, & Volkmar, 2003; Maestro, Filippo, Cavallaro, Pei, & al., 2002; Tantam, Holmes, & Cordess, 1993; Klin, Jones, Schultz, & Volkmar, 2003; Dawson, Meltzoff, Osterling, Rinaldi, & Brown, 1998). Some autistics even seem to find eye contact aversive (Hutt & Ounsted, 1966). In addition, delays in attention shifting have been observed experimentally (Casey, Gordon, Manheim, & Rumsey, 1993; Wainwright-Sharp & Bryson, 1993; Goldberg et al., 2002; Landry & Bryson, 2004). This deficit has been linked to cerebellar abnormalities (Harris, Courchesne, Townsend, Carper, & Lord, 1999).

In a previous modeling study (Triesch, Teuscher, et al., 2006) using a simpler model we demonstrated that a reduced saliency of the caregiver's face and/or delayed attention shifting impaired or abolished the emergence of gaze following. Here we tested the influence

of reduced caregiver saliency on the current model's learning behavior. In addition, autism has been associated with deficits in the mirror system (Dapretto et al., 2006). An obvious question is whether an "autistic" version of the new model would or would not develop a population of mirror neurons for gaze following. To this end we studied the following situations. First, we systematically reduced the caregiver saliency to model disinterest in faces as seen in autism. At the extreme, the caregiver's face was aversive. Second, we introduced varying delays in attention shifting. To this end, the model's decisions to shift gaze to a new location were delayed by 1–3 time steps.

When the saliency of the caregiver was reduced, gaze following emerged only slowly or not at all. This is illustrated in Fig. 7, which shows the development of gaze following ability of the model for various values of the caregiver saliency. We test the gaze following ability of the model every 100,000 training steps in a simulated experiment modelled after Corkum and Moore (1995). In these experiments, the caregiver turns to look at one side of the room ($\pm 80°$) in the absence of any visual targets, after making sure that the infant is looking at the caregiver. It is measured whether the infant turns to look to the correct side (score is +1), the opposite side (score is -1), or does not turn at all (score is 0) within six seconds after the caregiver's head turn (12 time steps). We average these scores across 100 repetitions to obtain the *gaze following score*, which is plotted in the figure for varying caregiver saliency. Lowering the caregiver saliency slowed down or even abolished the emergence of gaze following.

Figure 7. `<< about here >>`

Fig. 8(left) shows that this altered development of gaze following is accompanied by a

lack of connectivity between the representations of the caregiver's head and eye orientation and the pre-motor area. We plot the sum of the absolute values of all weights from layers **h** and **e** to layer **m** at the end of training as a function of the caregiver saliency. A dramatic decrease of the connectivity can be observed as the saliency is lowered to zero.

Figure 8. `<< about here >>`

This effect can be further corroborated by delayed attention shifting as show in the right part of the figure. Here we plot the absolute strength of the connectivity from layers **h** and **e** to layer **m** as a function of the amount of delay in attention shifting for two different values of the caregiver saliency. The connections become weaker with longer delays in attention shifting, although the effect is not as dramatic as that of reducing the caregiver saliency.

4 Discussion

We have presented a computational model of the emergence of gaze following based on biologically plausible reinforcement learning mechanisms. Despite its simplicity, the model seems to explain a large number of findings about the emergence of gaze following in human infants, some of which will be published elsewhere. These include the progression in expertise when following gaze to targets in different locations, the improving ability to ignore distractor objects, and the changing utilization of head pose and eye cues for gaze following (Butterworth & Jarrett, 1991; Corkum & Moore, 1995; Caron et al., 2002). A number of other models of the emergence of gaze following have been proposed in the past (Matsuda & Omori, 2001; Scassellati, 2002; Nagai et al., 2003; Lau & Triesch, 2004;

Hoffman et al., 2006), but, to the best of our knowledge, none of them accounts for the full range of experimental findings mentioned above. It should be noted, however, that in contrast to some of these (Scassellati, 2002; Nagai et al., 2003), our model has not yet been demonstrated on a real robot, which is a significant simplification. A general discussion of the relative merits of using robotic vs. simulation models has been given previously (Jasso & Triesch, 2005).

Our model was designed to offer a simple and parsimonious account of the complicated sequence of behavior patterns observed in the development of gaze following abilities in human infants, which will be discussed in depth in a forthcoming publication. Only after the model was completed, we realized that the representation in the model's pre-motor area shares important properties with the mirror neuron system in primates. The pre-motor representations in our model are different from the mirror neurons that have been reported in monkeys so far in the sense that they are not concerned with manual or oral motor acts or facial expressions but with gaze shifts. Thus, the model predicts the existence of a new class of mirror neurons specific to looking behaviors. If such a class of mirror neurons is found, it will lend support to our model.

This raises the important question of where in the brains of monkeys (or humans) one should look for such neurons. Electrophysiological and brain imaging studies suggest some tentative answers. Area F5, where the first mirror neurons (for grasping) were reported is an obvious candidate. More generally, the predicted class of mirror neurons might reside in any area intermediate between the superior temporal sulcus (STS), where head and gaze direction sensitive neurons are found, and eye movement related areas such as the frontal eye fields (FEF). Some of the predicted mirror neurons may also be present *inside* the FEF.

The model also predicts that this area should receive direct or indirect input from a visual saliency map. It is also conceivable that neurons with similar mirror properties can be found in the superior colliculus (SC). On the one hand, the SC receives inputs from the STS and represents salient stimuli in a retinotopic coordinate frame. On the other hand, the SC is contributing to the control of eye movements.

In our model, the mirror neurons emerge in two (not necessarily successive) steps. First, the model learns to perform certain motor acts in the appropriate situations. Concretely, it learns to map the discovery of visually salient stimuli in certain locations to gaze shifts to those locations. This corresponds to learning the appropriate pattern of weights between the representation of visually salient stimuli in layer **s** to the layer **m** of pre-motor neurons that encode the intention to make gaze shifts to specific locations. Second, the model learns to associate representations of the looking behavior of other agents to appropriate pre-motor neurons. This corresponds to learning the appropriate pattern of weights between the representation of the other agent's gaze direction in layer **h** and **e** to the same layer of pre-motor neurons **m**, thereby establishing an alternative pathway for activating neurons in this layer. This process is purely driven by the desire to maximize rewards, which, in this case, are obtained for looking at interesting visual stimuli. Thus, the gaze following behavior is learned because the gaze shift of another agent indicates that it is rewarding to perform the same gaze shift, i.e. to look at the same location. This leads to a number of implications for the emergence of imitation and the mirror neuron system.

4.1 Implications for our Understanding of Imitation

In the context of theories of imitation, our account of the emergence of gaze following can be considered a simple associative learning account of a response facilitation. It does not start with any mechanism (or "module") for, say, matching other's bodies to one's own. On the contrary, our model has a generic reinforcement learning architecture. Nevertheless, it *acquires* the ability to map others' motor acts onto its own behaviors. This finding may be of interest for the question of the development of higher imitative behaviors. While it may be that specific mappings from other bodies to one's own body are present at birth (e.g., Meltzoff, 2005), we have shown that such mappings can also result from generic reinforcement learning mechanisms in a parsimonious fashion.

4.2 Implications for our Understanding the Mirror Neuron System

Our model also has implications for the question whether mirror neurons are innate or whether they acquire their special properties through a learning process. So far, there have been no studies investigating to what extent mirror neurons may already be present in the infantile brain, or what kinds of experiences and interactions with the environment are necessary and sufficient for their emergence. However, a number of theoretical accounts for the emergence of mirror neurons through learning processes have been offered previously (Heyes, 2001; Oztop & Arbib, 2002; Keysers & Perrett, 2004; Brass & Heyes, 2005; Weber, Wermter, & Elshaw, 2006; Jones, 2006; Metta, Sandini, Natale, Craighero, & Fadiga, 2006; Oztop, Kawato, & Arbib, 2006). These accounts are based on Hebbian learning in the context of self-observation and/or being imitated by other agents, not reinforcement learning. Thus, the mechanism of the emergence of mirror neurons for gaze following in

our model is distinct from and complementary to these accounts. This raises an interesting question: could the same reinforcement learning mechanism also contribute to the emergence of other kinds of mirror neurons? For the "classic" mirror neurons concerned with grasping, we find it plausible that there are situations where observing an agent grasp an object (e.g., a food item grasped by the mother in order to eat it) may predict a reward if the same action is attempted (grasping a second food item from the same source in order to also consume it). Such situations may be sufficient for the emergence of mirror neurons for grasping. More generally, our reinforcement learning explanation may be applicable in many instances where imitation (of various forms) and a corresponding set of mirror neurons is observed. This theory predicts a very close connection between mirror neurons and imitative behaviors. It should be noted, however, that so far there is only little data on the involvement of mirror neurons in imitation.

To resolve this issue, it will be best to study the emergence of imitation and the mirror system longitudinally. If the appearance of certain imitative behaviors (such as gaze following or manipulative movements) during an individual's development turns out to coincide with the appearance of mirror neurons for these behaviors, this would be consistent with our hypothesis. More critically, however, our model predicts that if an animal were raised without the opportunity to ever observe a specific action performed by other animals, then no mirror neurons specific to this action should develop (see also Meltzoff, 2005 for a different but related proposal). Furthermore, if an animal were raised in an environment where it is rewarding to perform a behavior A whenever another agent performs a different behavior B, we would expect the emergence of "generalized" mirror neurons that respond to the animal performing A or to the observation of another animal performing B. Note

that such a "generalized" mirror neuron could not develop simply through self-observation or being imitated.

4.3 Implications for Our Understanding of Autism Spectrum Disorder

We have used the model to explore the potential origins of a putative dysfunction of the mirror system in autism (Oberman et al., 2005; Dapretto et al., 2006) that may be related to many of the behavioral problems associated with autism. It has recently been observed that autism is associated with a reduced thickness of the cortex in supposed mirror areas as well as in the STS (Hadjikhani, Joseph., Snyder, & Tager-Flusberg, 2006). What is unclear, however, is whether this is the cause or the result of (some of) the behavioral problems in autism. Our model shows that behavioral problems in gaze following can be explained by an initial disinterest in or aversion to faces that could be due to abnormalities in the amygdala, e.g. (R. Schultz, 2005). This can alter the developmental trajectory such that neither gaze following nor a corresponding set of mirror neurons will emerge. More precisely, areas representing the head and gaze orientation of faces such as the STS will not develop strong connections to pre-motor areas involved in the planning of gaze shifts (such as the FEF). This very specific reduction in connectivity is the anatomical reflection of the failure to develop mirror neurons. This may also cause a relative thinning of these areas because they will have fewer efferent projections and fewer afferent inputs compared to the normally developing cortex. At the same time, a reduction of the size of the area of STS representing head and eye gaze can be expected due to fewer gaze shifts to the caregiver because of an early disinterest in social stimuli.

4.4 Conclusion

Over the last few years, a number of theoretical accounts of the ontogeny of gaze following on the one hand and imitation and mirror neurons on the other hand have been proposed. Our model suggests that there may be a much closer link between these different developments than was previously thought. Further empirical and modeling work must aim to better understand the relationships between attention sharing and imitation skills, their joint development, and their neural basis.

Acknowledgements

This work was done as part of the MESA project (Modeling the Emergence of Shared Attention) at the University of California, San Diego (http://mesa.ucsd.edu). We thank the members of the MESA team for their continuing collaboration. We also thank Jaime Pineda and Garrison Cottrell for fruitful discussions and two anonymous reviewers for comments on an earlier draft. This work was supported by the National Science Foundation under grant SES-0527756. JT acknowledges support from the Hertie foundation and the European Union under grant MEXT-CT-2006-042484.

References

Adrien, J. L., Lenoir, P., Martineau, J., Perrot, A., & al. et. (1993). 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.

Baron-Cohen, S. (1995). *Mindblindness: an essay on autism and theory of mind*. Cambridge, MA: A Bradford Book, The MIT Press.

Brass, M., & Heyes, C. (2005). Imitation: is cognitive neuroscience solving the correspondence problem? *Trends in Cognitive Sciences*, *9*(10), 489–495.

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

Carlson, E., & Triesch, J. (2004). A computational model of the emergence of gaze following. In H. Bowman & C. Labiouse (Eds.), *Connectionist models of cognition and perception II* (pp. 105–114). Singapore: World Scientific.

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

Casey, B., Gordon, C., Manheim, G., & Rumsey, J. (1993). Dysfunctinoal attention in autistic savants. *Journal of Clinical and Experimental Neuropsychology*, *15*(6), 933–946.

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

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

Dapretto, M., Davies, M., Pfeifer, J., Scott, A., Sigman, M., Bookheimer, S., et al. (2006). Understanding emotions in others: mirror neuron dysfunction in children with autism spectrum disorders. *Nature Neuroscience*, *9*(1), 28–30.

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

Dawson, G., Toth, K., Abbott, R., Osterling, J., Munson, J., Estes, A., et al. (2004). Early social attention impairments in autism: Social orienting, joint attention, and attention to distress. *Developmental Psychology*, *40*(2), 271–283.

Dayan, P., & Abbott, L. F. (2001). *Theoretical Neuroscience*. Cambridge, MA: MIT Press.

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

D'Entremont, B., Hains, S., & Muir, D. (1997). A demonstration of gaze following in 3- to 6-month-olds. *Infant Behavior and Development*, *20*(4), 569–572.

Farroni, T., Massaccesi, S., Pividori, D., & Johnson, M. H. (2004). Gaze following in newborns. *Infancy*, *5*(1), 39–60.

Fasel, I., Deák, G. O., Triesch, J., & Movellan, J. (2002). Combining embodied models and empirical research for understanding the development of shared attention. In *Intl. Conf. on Development and Learning (ICDL)*. Cambridge, MA: IEEE Computer Society Press.

Flom, R., Deák, G., Phill, C. G., & Pick, A. D. (2003). Nine-month-olds' shared visual attention as a function of gesture and object location. *Infant Behavior and Development*, *27*, 181–194.

Gallese, V., Keysers, C., & Rizzolatti, G. (2004). A unifying view of teh basis of social cognition. *Trends in Cognitive Sciences*, *8*(9), 396–403.

Goldberg, M., Lasker, A., Zee, D., Garth, E., Tien, A., & Landa, R. (2002). Deficits in the initiation of eye movements in the absence of a visual target in adolescents with high functioning autism. *Neuropsychologica*, *40*(12), 2039–2049.

Hadjikhani, N., Joseph., R., Snyder, J., & Tager-Flusberg, H. (2006). Anatomical differences in the mirror neuron system and social cognition network in autism. *Cerebral Cortex*, *16*, 1276–1282.

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

Harris, N., Courchesne, E., Townsend, J., Carper, R., & Lord, C. (1999). Neuroanatomic contributions to slowed orienting of attention in children with autism. *Cognitive Brain Research*, *8*(1), 61–71.

Hayhoe, M., Land, M., & Shrivastava, A. (1999). Coordination of eye and hand movements in a normal environment. *Invest. Opthalmol. & Vision Science*, *40*.

Heyes, C. (2001). Causes and consequences of imitation. *Trends in Cognitive Sciences*, *5*(6), 253–261.

Hoffman, M., Grimes, D., Shon, A., & Rao, R. (2006). A probabilistic model of gaze imitation and shared attention. *Neural Networks*, *19*, 299–310.

Hutt, C., & Ounsted, C. (1966). The biological significance of gaze aversion with particular reference to the syndrome of infantile autism. *Behavioral Science*, *11*(5), 346–356.

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J., & Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. *PLOS Biology*, *3*, 529–535.

Iacoboni, M., Woods, R., Brass, M., Bekkering, H., Mazziotta, J., & Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science*, *286*, 2526–2528.

Jasso, H., & Triesch, J. (2005). A virtual reality platform for modeling cognitive development. In G. Palm & S. Wermter (Eds.), *Biomimetic neural learning for intelligent robots.* Berlin / Heidelberg: Springer.

Jasso, H., & Triesch, J. (2006). Using eye direction cues for gaze following — a developmental model. In *Proc. of the Int. Conf. on Development and Learning (ICDL 2006).* Bloomington, IN: Indiana University.

Jasso, H., Triesch, J., Teuscher, C., & Deák, G. (2006). A reinforcement learning model explains the development of gaze following. In *Int. Conf. on Cognitive Modeling (ICCM).* Bagnaria Arsa, Italy: Edizioni Goliardiche.

Jenkins, R., Beaver, J., & Calder, A. (2006). I thought you were looking at me: Direction-specific aftereffects in gaze perception. *Psychological Science*, *17*(6), 506–513.

Jones, S. S. (2006). Infants learn to imitate by being imitated. In *Proc. Int. Conf. on Development and Learning (ICDL).* Bloomington, IN: Indiana University.

Keysers, C., & Perrett, D. (2004). Demystifying social cognition: a Hebbian perspective. *Trends in Cognitive Sciences*, *8*(11), 501–507.

Klin, A., Jones, W., Schultz, R., & Volkmar, F. (2003). The enactive mind, or from actions to cognition: lessons from autism. *Phil. Trans. R. Soc. Lond. B*, *358*, 345–360.

Kohler, E., Keysers, C., Umiltà, M. A., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: Action representation in mirror neurons. *Science*, *297*, 846–848.

Kumashiro, M., Ishibashi, H., Uchiyama, Y., Itakura, S., Mrata, A., & Iriki, A. (2003). Natural imitation induced by joint attention in japanese monkeys. *Int. J. Psychophysiology*, *50*, 81–99.

Landry, R., & Bryson, S. (2004). Impaired disengagement of attention in young children with autism. *Journal of Child Psychology & Psychiatry*, *45*(6), 1115–1122.

Lau, B., & Triesch, J. (2004). Learning gaze following in space: a computational model. In *Intl. Conf. on Development and Learning (ICDL).* La Jolla, CA: The Salk Institute for Biological Studies.

Leslie, A. M. (1987). Pretense and representation — the origins of theory of mind. *Psychological Review*, *94*(4), 412–426.

Maestro, S., Filippo, M., Cavallaro, M., Pei, F., & al. et. (2002). 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.

Matsuda, G., & Omori, T. (2001). Learning of joint visual attention by reinforcement learning. In E. M. Altmann & A. Cleeremans (Eds.), *Proc. Int. Conf. on Cognitive Modeling (ICCM)* (pp. 157–162). Mahwah, NJ, US: Lawrence Erlbaum Associates.

Meltzoff, A. N. (2005). Imitation and other minds: The "like me" hypothesis. In S. Hurley & N. Chater (Eds.), *Perspectives on imitation: From neuroscience to social science* (pp. 55–77). Cambridge, MA: MIT Press.

Metta, G., Sandini, G., Natale, L., Craighero, L., & Fadiga, L. (2006). Understanding mirror neurons: a bio-robotic approach. *Interaction Studies*, *in press.*

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

Nagai, Y. (2005). Joint attention development in infant-like robot based on head movement imitation. In *Proc. Third Int. Symposium on Imitation in Animals and Artifacts (AISB'05)* (pp. 87–96). Brighton, UK: The Society for the Study of Artificial Intelligence and the Simulation of Behaviour.

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

Oberman, L., Hubbard, E., McCleery, J., Altschuler, E., Ramachandran, V., & Pineda, J. (2005). EEG evidence for mirror neuron dysfunction in autism spectrum disorders. *Cognitive Brain Research*, *24*, 190–198.

Osterling, J., & Dawson, G. (1994). Early recognition of children with autism: a study of first birthday home video tapes. *Journal of Autism and Developmental Disorders*, *24*, 247–257.

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

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

Pobric, G., & Hamilton, A. d. C. (2006). Action understanding requires the left inferior frontal cortex. *Current Biology*, *16*, 524–529.

Rizzolatti, G. (2005). The mirror neuron system and its function in humans. *Anat. Embryol.*, *210*, 419–421.

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annu. Rev. Neurosci.*, *27*, 169–192.

Sai, F., & Bushnell, W. R. (1989). The perception of faces in different poses by 1-month-olds. *British Journal of Developmental Psychology*, *6*, 35–41.

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

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

Schaal, S. (1999). Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences*, *3*, 233–242.

Schultz, R. (2005). Developmental deficits in social perception in autism: the role of the amygdala and fusiform face area. *Int. J. of Developmental Neuroscience*, *23*, 125–141.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1599.

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

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

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

Teuscher, C., & Triesch, J. (2006). *To each his own: The caregiver's role in a computational model of gaze following.* Neurocomputing, to appear.

Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge, MA: Harvard Univ. Press.

Triesch, J., Jasso, H., & Deák, G. (2006). Emergence of mirror neurons in a model of gaze following. In *Proc. of the Int. Conf. on Development and Learning (ICDL 2006).* Bloomington, IN: Indiana University.

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

Umiltà, M., Kohler, E., Gallese, V., Fogassi, L., L., F., Keysers, C., et al. (2001). "I know what you are doing:" a neurophysiological study. *Neuron*, *31*, 91–101.

Wainwright-Sharp, J., & Bryson, S. (1993). Visual orienting deficits in high-functioning people with autism. *Journal of Autism and Developmental Disorders*, *13*(1), 1–13.

Weber, C., Wermter, S., & Elshaw, M. (2006). A hybrid generative and predictive model of the motor cortex. *Neural Networks*, *19*(4), 339-53.

Williams, J., Whiten, A., & Singh, T. (2004). A systematic review of action imitation in autistic spectrum disorder. *Journal of Autism and Developmental Disorders*, *34*(3), 285–299.

Table and Figure Captions

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

Figure 1 Learning environment: an infant and a caregiver interact with a number of objects. Note that the caregiver's direction of gaze is not always perfectly aligned with the orientation of the caregiver's head. The region that is not visible for the infant, because it is outside of the infant's current field of view, is drawn hatched.

Figure 2 Actor-critic reinforcement learning model for learning gaze following. A number of visual areas process the visual input in terms of a *saliency map* **s** and representations of the *caregiver head direction* **h**, and the *caregiver eye direction* **e**. These visual representations serve as input to the critic, who learns to make reward predictions, and the actor, who is composed of a pre-motor area and an action selection mechanism.

Figure 3 Learned connection weights from layers **h** and **e** to the pre-motor layer **m**. Each pre-motor unit is usually strongly activated by just one or two units in **h** and **e** each. Conversely, however, a single unit in **h** and **e** will activate several units in **m**. The efferent weights of marked units h_7 and e_{11} are illustrated in Fig. 4. The properties of the marked pre-motor units A and B are studied in Figs. 5 and 6.

Figure 4 Illustration of connection weights from eye-direction unit e_{11} in layer **e** and head direction unit h_7 in layer **h** to the pre-motor layer **m** (compare Fig. 3). The units have developed strengthened connections to pre-motor units representing the appropriate line-of-sight of the caregiver.

Figure 5 Illustration of motor properties of two pre-motor units in layer **m** (compare Fig. 3). We compare the distributions of activations of the two neurons when a gaze shift to their target locations is made vs. when a gaze shift to a different location is made. The sensory properties of the same two units are illustrated in Fig. 6.

Figure 6 Sensory properties of the two marked pre-motor units from Fig. 3. The unit activations are shown for a number of stimulation conditions while the infant is looking at the caregiver. Both units become more activated when the caregiver is looking in their direction. See text for details.

Figure 7 Gaze following performance as a function of training time for different caregiver saliencies. For low caregiver saliency gaze following will emerge only very slowly or not at all. Error bars represent standard error of the mean (10 simulations).

Figure 8 Sum of absolute values of weights from layers **h** and **e** to layer **m** after learning as a function of caregiver saliency (left) and latency of attention shifting (right). Reducing the caregiver saliency dramatically reduces the strength of connections from the **h** and **e** layers to the **m** layer as shown left. A deficit in attention shifting leads to a similar albeit smaller reduction in weight strength as shown for two values of the caregiver saliency on the right. Error bars represent standard error of the mean (10 simulations).

Symbol	**Explanation**	**Range**	**Default**
	Environment Parameters		
Φ_I	Infant's saliency	$(-\infty, \infty)$	2.0
Φ_C	Caregiver's saliency	$(-\infty, \infty)$	2.0
$\bar{\Phi}_o$	Average object saliency	$(-\infty, \infty)$	1.0
$\bar{N}_o$	Average number of objects	$[0, \infty)$	4
σ_o	Object placement spread	$[0, \infty)$	0.5 m
$\bar{T}_{\text{present}}$	Average caregiver presence interval	$[0, \infty)$	120
$\bar{T}_{\text{absent}}$	Average caregiver absence interval	$[0, \infty)$	120
$\bar{T}_{\text{objects}}$	Average object replacement interval	$[0, \infty)$	10
σ_C	Standard deviation of caregiver's head orientation	$[0, \infty)$	5°
	Infant Visual System Parameters		
σ_F	Foveation range	$[0°, \infty)$	180°
τ_H	Habituation rate	$[0, \infty)$	1.2
α_H	Target of habituation	$[1, \infty)$	1.0
d	Memory decay factor	[0,1]	0.5
	Infant Learning Parameters		
η	Learning rate for synaptic weights $\mathbf{m}$ and $\mathbf{M}$	$[0, \infty)$	0.005
γ	Discount factor for future rewards	$[0, 1)$	0.2
β	Inverse temperature for softmax action selection	$[0, \infty)$	100

Table 1

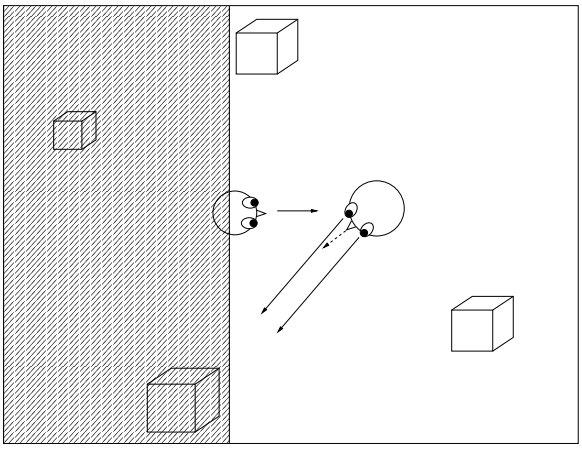

Figure 1

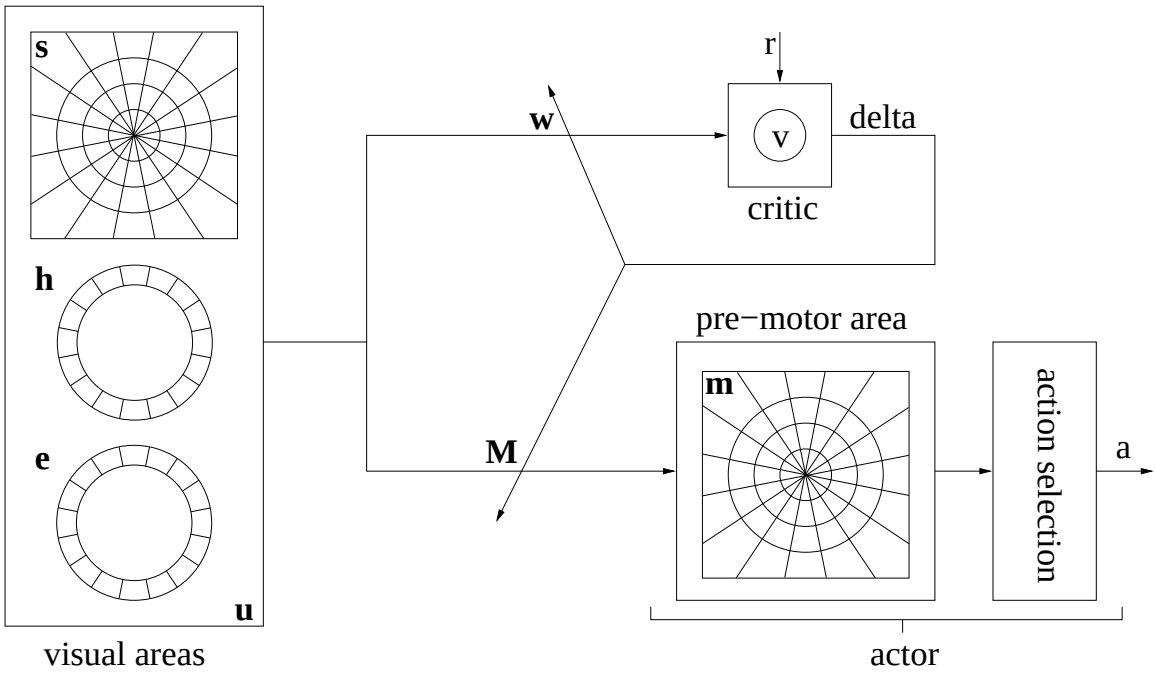
s
h
e
u
visual areas
r
w
v
delta
critic
pre−motor area
m
M
action selection
a
actor

Figure 2

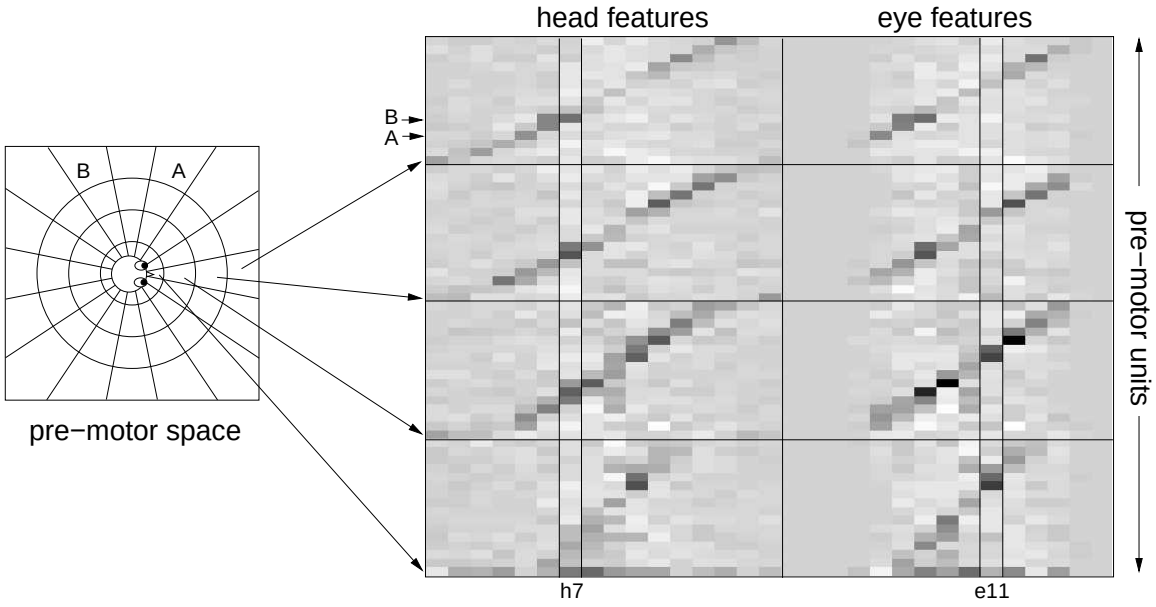
head features
eye features
B
A
B
A
pre−motor space
pre−motor units
h7
e11

Figure 3

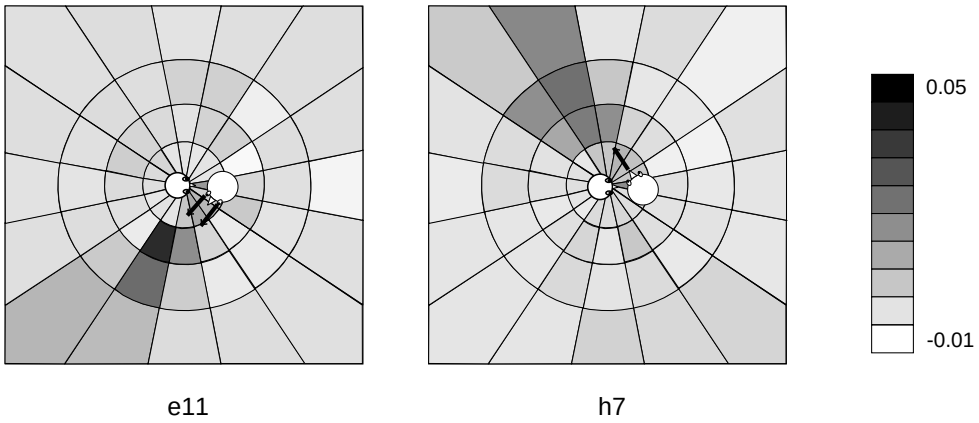

Figure 4

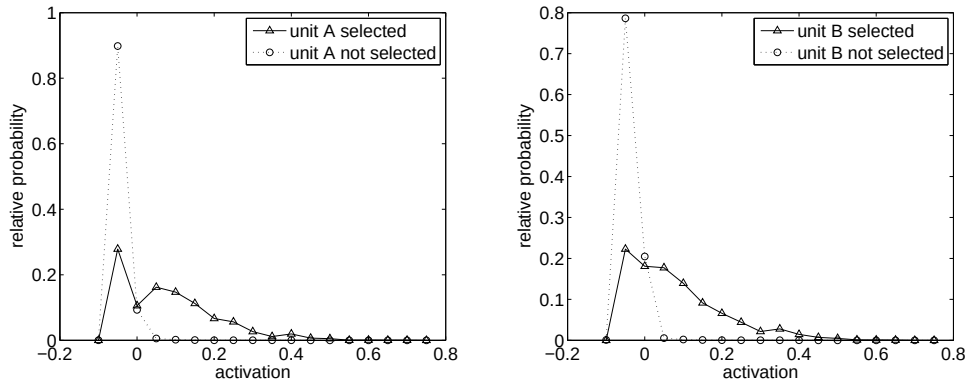
unit A selected
unit A not selected
relative probability
activation
unit B selected
unit B not selected
relative probability
activation

Figure 5

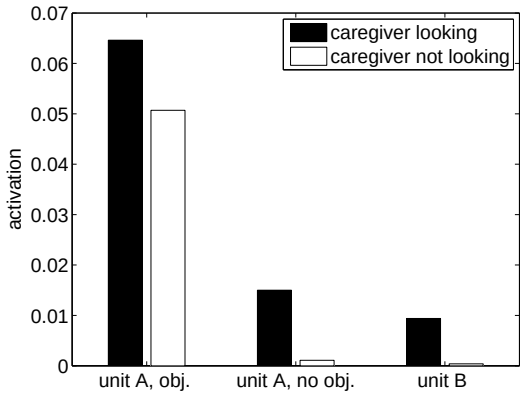
caregiver looking
caregiver not looking
activation
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
unit A, obj.
unit A, no obj.
unit B

Figure 6

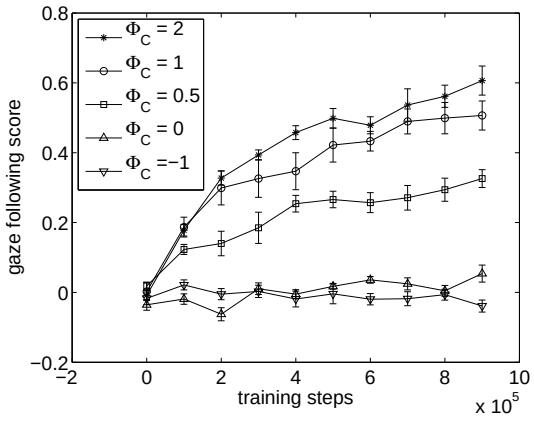
0.8
0.6
0.4
0.2
0
−0.2
−2
0
2
4
6
8
10
gaze following score
training steps
x 10^5
Φ_C = 2
Φ_C = 1
Φ_C = 0.5
Φ_C = 0
Φ_C =−1

Figure 7

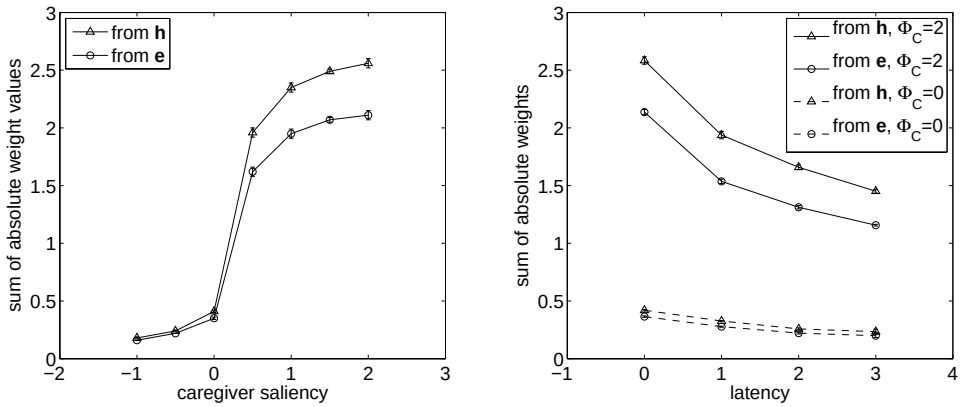
from **h**
from **e**
sum of absolute weight values
caregiver saliency
from **h**, Φ_C=2
from **e**, Φ_C=2
from **h**, Φ_C=0
from **e**, Φ_C=0
sum of absolute weights
latency

Figure 8

Author Biographies

Jochen Triesch studied physics and received his Diploma degree in 1994 and his Ph.D. degree in 1999, both from the University of Bochum, Germany. After two years as a post-doctoral fellow at the University of Rochester, USA, he joined the faculty of the University of California San Diego in 2001 as an Assistant Professor of Cognitive Science. Since 2005 he is also a Fellow of the Frankfurt Institute for Advanced Studies in Frankfurt am Main, Germany. Email: `triesch@fias.uni-frankfurt.de`. Address: Frankfurt Institute for Advanced Studies, Max-von-Laue-Str. 1, D-60438 Frankfurt am Main, Germany.

Hector Jasso is a Ph.D. Candidate in Computer Science at the University of California, San Diego. He received his B.Sc. in Computer Science at the Instituto Tecnologico y de Estudios Superiores de Monterrey (ITESM), Mexico, in 1988, and an M.Sc. in Information Technology - Knowledge Based Systems, at the University of Edinburgh, Scotland, in 1991. Email: hjasso@cs.ucsd.edu. Address: Department of Computer Science and Engineering, University of California, San Diego, 9500 Gilman Drive, Mail Code 0404, La Jolla, CA 92093-0404, USA.

Gedeon O. Deák is an Associate Professor of Cognitive Science and Human Development at the University of California at San Diego, where he has been on the faculty since 1999. He received his BA from Vassar College (1990) and his Ph.D. from the Institute of Child Development at the University of Minnesota (1995). He was on the faculty at Vanderbilt University from 1995 to 1999. As director of the Cognitive Development Laboratory (`http://www.cogsci.ucsd.edu/~deak/cdlab/`), he studies social-cognitive development,

language, and problem solving. Email: deak@cogsci.ucsd.edu. Address: Dept. of Cognitive Science, 9500 Gilman Drive, La Jolla, CA 92093-0515, USA.