Neural network models of cognitive development in infancy

DISSERTATION
zur
Erlangung des Grades
Doktor der Naturwissenschaften

vorgelegt beim Fachbereich Physik
der Goethe-Universität Frankfurt am Main

von
Arthur Franz
Dipl. Phys.

Frankfurt am Main (2010)

Vom Fachbereich Physik der Goethe-Universität Frankfurt am Main
als Dissertation angenommen.

Dekan: Prof. Dr. Dirk H. Rischke

1. Gutachter: Prof. Dr. Jochen Triesch

2. Gutachter: Prof. Dr. Christoph von der Malsburg

Acknowledgments

It is a pleasure to thank all those people who have contributed to this thesis.

I owe my deepest gratitude to my first mentor, Jochen Triesch, without whom this thesis would not have been possible. On the one hand he supported me with suggestions, guidance and feedback at any time and without him I would have been lost being a young researcher. On the other hand, he gave me all the freedom I need to develop my own thoughts and research directions.

It is an honor for me to have had Christoph von der Malsburg as my second mentor. Maybe without knowing it himself, but by being an example, he taught me not to lose the ambition for the biggest questions in science especially in the days of tedious, everyday scientific practice.

I would like to show my gratitude to Thorsten Kolling and Monika Knopf who put much effort in setting up a collaboration between FIAS and the developmental psychology group and giving me first hand experience with infant research.

I am grateful to my colleagues to support me and to create a good working environment. Specifically, I would like to thank Prashant Joshi for the fun scientific discussions, feedback on my publications and a healthy work-life balance. Special thanks goes to Andreea Lazar who supported me in setting up her model that the project in Chapter 5 is based on.

Любимые родители! Я хочу вам выразить огромную благодарность, так как без вашей любви, поддержки и оптимизма, я никогда бы не достиг таких высот и этой диссертацией бы и не пахло. Это честь, быть вашим сыном.

Contents

List of Figures

List of abbreviations

AI	Artificial Intelligence
IP	Intrinsic Plasticity
LLLR	Left-Left-Left-Right sequence
LLR	Left-Left-Right sequence
LR	Left-Right alternating sequence
LTLB	Left-Top-Left-Bottom (pivot) sequence
MRM	Minimal Required Memory
RL	Reinforcement Learning
RLA	Reinforcement Learning Architecture
RNN	Recurrent Neural Network
SN	Synaptic Normalization
STDP	Spike Timing Dependent Plasticity

Abstract

This thesis investigates the development of early cognition in infancy using neural network models. Fundamental events in visual perception such as caused motion, occlusion, object permanence, tracking of moving objects behind occluders, object unity perception and sequence learning are modeled in a unifying computational framework while staying close to experimental data in developmental psychology of infancy.

In the first project, the development of causality and occlusion perception in infancy is modeled using a simple, three-layered, recurrent network trained with error backpropagation to predict future inputs (Elman network). The model unifies two infant studies on causality and occlusion perception. Subsequently, in the second project, the established framework is extended to a larger prediction network that models the development of object unity, object permanence and occlusion perception in infancy. It is shown that these different phenomena can be unified into a single theoretical framework thereby explaining experimental data from 14 infant studies. The framework shows that these developmental phenomena can be explained by accurately representing and predicting statistical regularities in the visual environment. The models assume (1) different neuronal populations processing different motion directions of visual stimuli in the visual cortex of the newborn infant which are supported by neuroscientific evidence and (2) available learning algorithms that are guided by the goal of predicting future events. Specifically, the models demonstrate that no innate force notions, motion analysis modules, common motion detectors, specific perceptual rules or abilities to "reason" about entities which have been widely postulated in the developmental literature are necessary for the explanation of the discussed phenomena.

Since the prediction of future events turned out to be fruitful for theoretical explanation of various developmental phenomena and a guideline for learning in infancy, the third model addresses the development of visual expectations themselves. A self-organising, fully recurrent neural network model that forms internal representations of input sequences and maps them onto eye movements is proposed. The reinforcement learning architecture (RLA) of the model learns to perform anticipatory eye movements as observed in a range of infant studies. The model suggests that the goal of maximizing the looking time at interesting stimuli guides infants' looking behavior thereby explaining the occurrence and development of anticipatory eye movements and reaction times. In contrast to classical neural network modelling approaches in the developmental literature, the model uses local learning rules and contains several biologically plausible elements like excitatory and inhibitory spiking neurons, spike-timing dependent plasticity (STDP), intrinsic plasticity (IP) and synaptic scaling. It it also novel from the technical point of view as it uses a dynamic recurrent reservoir shaped by various plasticity mechanisms and combines it with reinforcement learning. The model accounts for twelve experimental studies and predicts among others anticipatory behavior for arbitrary sequences and facilitated reacquisition of already learned sequences.

All models emphasize the development of the perception of the discussed phenomena thereby addressing the questions of how and why this developmental change takes place - questions that are difficult to be assessed experimentally. Despite the diversity of the discussed phenomena all three projects rely on the same principle: the prediction of future events. This principle suggests that cognitive development in infancy may largely be guided by building internal models and representations of the visual environment and using those models to predict its future development.

Chapter 1

Introduction

> *"Studying babies is like studying the Big Bang of cognition. It is where you can see the building blocks of cognition coming into being and taking shape, and you start to see the structures that come to make up our adult cognitive universe." (David Rakison)*

In the past decades many fields of interdisciplinary research have developed motivated by the attempt to understand human cognition. Classically, artificial intelligence (AI) was mainly dominated by the attempt to make a machine sense and act intelligently, while the research was largely guided by an intuitive notion of "intelligence". Specifically, the research was not focused on understanding humans but rather motivated by a poorly understood, introspective notion of what it means to see, think and act.

With time, people recognized that this kind of research stands on loose grounds and directed their gaze at actually studying humans, the results of psychology and neuroscience. The field of cognitive science emerged - an interdisciplinary field consisting of psychology, neuroscience, artificial intelligence, mathematics and philosophy. Additionally, researchers coming from the "hard" sciences recognized the lack of theory and a confusingly large amount of empirical data that still permeate the biological and psychological sciences. Subsequently, many theoretical and computational models of psychological and biological phenomena emerged with time. This thesis constitutes an effort to contribute to this line of work in computational developmental psychology.

1.1 Motivation: why study infants?

This thesis consists of a line of thought mainly motivated by the following questions: how do infants get started with understanding the world? Specifically, how do infants learn to see? How can we use that knowledge to teach a machine to do the same? Classically, AI researchers have tried to build in our knowledge about the world by hand, by ceaselessly typing in millions of common sense knowledge statements into the machine in order to make it know and to enable it to perform the same kinds of inferences and tasks that even three-year-olds master without effort. These projects turned out to be unfeasible and made us understand the amount of assumptions and knowledge that each of us has acquired since our birth.

Nowadays, most experts agree that the only way to make a machine know what we know is to make it *learn* in a similar way that we do. Although this seems to constitute a promising way of looking at the problem, it turns out that people use a huge set of intrinsic assumptions even when they learn new things. It seems that the adult human is too complex a system to get started although research on adult humans is certainly necessary and pragmatically meaningful. Therefore, if we want to study the Big Bang of human cognition the human infant seems like a good starting point.

1.2 The origins of knowledge

1.2.1 Innate or learned?

Accepting that the initial, most important pieces of knowledge about the world are acquired in infancy begs the question about what that knowledge is. Additionally, we have to ask whether this knowledge is given at birth, maturates during development (nativism) or whether it is learned (empiricism). This nature - nurture debate can be traced back to the empiricists John Locke and David Hume and the reactions to their views. Although considered as obsolete by many researchers, this debate still goes on, see e.g. Spencer et al. (2009).

For example, one of the most prominent representatives of current nativism is Elizabeth S. Spelke who advocated the "core knowledge" hypothesis (Spelke and Kinzler, 2007). As for the domain of knowledge about naïve physics, infants are suggested to be endowed with innate knowledge about "cohesion (objects move as connected, bounded units), continuity (objects move on connected, unobstructed paths), and contact (objects affect one another's motion if and only if they touch)" (Spelke, 1994). Given that knowledge, it is conceivable that infants are well equipped to start acquiring, maybe learning, even more about the objects they are surrounded with.

Despite the convenience of nativist suggestions, for the present work, these sorts of views are unsatisfactory for the following reasons. First, for any of the infant's skills, whether learned or innate, we strive to understand how it came about and merely postulating that it is innate does not satisfy our intellectual demand. Second, it is unfortunately not rare that empirical research on infants is used to extrapolate beyond the data to serve nativist arguments. For example, merely the fact that infants have some cognitive ability as early as at the age of four months does not prove that it is innate (i.e. given from birth), see Spencer et al. (2009) for a detailed critique. Finally, from the point of view of a theoretician attempting to model development, (speculative) innate skills represent additional assumptions on his theory which he desires to avoid. On the contrary, it seems attractive to ask how we can build a theory of infant development in which various abilities are acquired through learning and following a similar development as observed in infants while making as few assumptions as possible. Only at the end of this long research effort, if we are able to state with some certainty that there is a small set of assumptions that are really necessary to get an infant started, only then it seems justified to postulate them as innate.

1.2.2 The point of debarkment

Given the considerations in the previous section and our decision to study how infants visually start to extract the relevant knowledge from the world, let us describe the newborn infant's impression of the world as did William James: "one great blooming, buzzing confusion" James (1890), p. 462. Imagine that all the infant has is this chaos of colorful pixels of light that fall on his retina, the ability to move the eyes and one or more powerful learning mechanisms. How far can we get from here? It is emphasized that it is by no means asserted that an infant actually is this tabula rasa. It just seems fruitful from a theoretician's stance to adopt this view as a starting point justified by pragmatic arguments given in the previous section.

Further it is necessary to investigate which capacities, what knowledge actually develops and when it develops under the constraint that it is conceivable that some learning mechanism may lead the infant from its starting point at birth to these pieces of knowledge. In other words, it seems instructive to look for pieces of knowledge that infants have been observed to have at a very young age, such as the first months of life.

At this point the theoretician is confronted with a wide array of possible things that infants could know and the difficulty to constrain his efforts to a small set of phenomena. From a pragmatic point of view though, it is useful to concentrate on visual scenes and phenomena that actually have been studied experimentally well enough to present a sufficiently rich data set for theoretical modeling. In the following we will review some of the relevant data serving as starting point of this thesis while also aiming at familiarizing the reader with methods in developmental psychology of infancy.

1.2.3 Object segmentation and unity

Continuing our heuristic approach, it seems reasonable for the newborn infant to start segmenting the visual scene into separate objects since it is not clear from

the chaos of colorful pixels where objects, including people, begin and end. One of the prevailing hypotheses known as *common fate* (Wertheimer, 1923) is that those parts of the image that move together also belong together as parts of the same object.

Of course it is not possible to ask preverbal infants whether they perceive an object as united or not. Children utter their first words only by the age of 12-18 months (Clark, 2004). Therefore, developmental psychologists have developed an array of methods in order to investigate what young infant might think. One of the prevailing methods is the so-called habituation paradigm. Infant habituation is normally studied within a framework derived from the pioneering work of Soviet physiologist Evgeni Sokolov and his colleagues (Vinogradova, 1975, Sokolov, 1963, 1975). Within this framework, infants are assumed to build (through learning) a neural or mental model of stimuli or stimulus events as these are repeatedly presented. When a stimulus is presented, it is compared to this neural model and discrepancies provide the basis for learning. This is often referred to as the comparator theory (Gilmore and Thomas, 2002). As learning progresses, discrepancies between the model and external events decrease, which result in a decrease of attention (i.e., there is a progressively smaller need to process the information as the internal model approximates such information). Crucially, it is assumed that stimuli that deviate from this internal model would elicit a relative increase in responding (a behavior referred to as dishabituation, release from habituation, or renewed responding). Indeed, a stimulus perceived as novel would require more processing than one perceived as familiar.

Renewed responding thus provides researchers with a unique opportunity to investigate internal representations in preverbal infants. What infants perceive as novel (inferred from renewed responding) given a set of habitual stimuli or events allows for inferences as to how they represent information. By carefully designing stimuli sets, researchers can systematically examine in which ways stimuli may be represented as distinct.

Kellman and Spelke (1983) did a classic study based on the habituation paradigm investigating whether 4-month-olds are able to perceive unity of objects. A rod

was moved behind a block as depicted in Fig. 1.1 and it was investigated whether infants perceive the ends of the rod as connected behind the block. The subjects were separated into two groups, one habituated with a complete rod moving behind the block, the "rod movement" group, and another habituated with only the upper rod piece moving while the lower rod piece remained stationary, the "baseline" group. The baseline was introduced in order to rule out alternative explanations based only on the way the involved objects look like. The infants were shown the habituation stimuli repeatedly and their looking time was measured until the mean looking time of three consecutive trials dropped below 50% of the mean looking time of the first three consecutive trials.

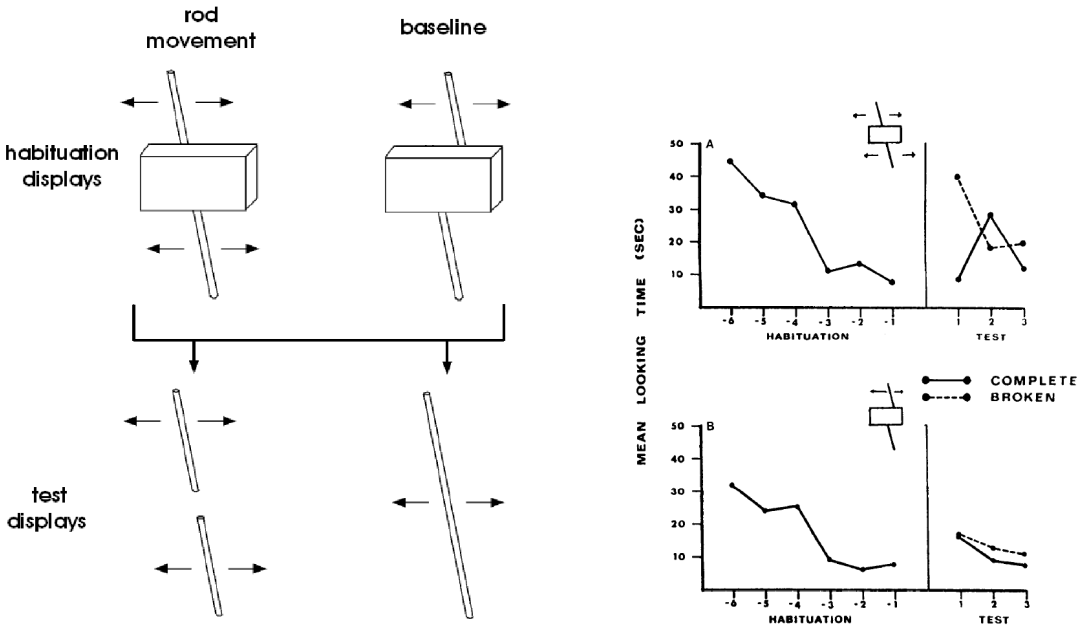

Figure 1.1: Displays, design and results in the study by Kellman and Spelke (1983). Left: Two groups of infants were habituated to the "rod movement" and "baseline" displays, respectively. Then they were tested with the "complete rod" and "broken rod displays appearing in an alternating fashion. The measurement of looking time indicated infants' interpretation of the habituation stimulus. Right: Looking times at the stimuli by infants (A) in the rod movement group and (B) in the baseline group. Note: For statistical analysis of significance the reader is referred to the original publication.

After habituation both infant groups were exposed to two test stimuli: the "complete rod" test and the "broken rod" test as depicted in Fig. 1.1, left. The test stimuli were shown in an alternating way, three times each, resulting in six test displays. For each test trial the looking time was measured again.

Fig. 1.1, right, shows that infants in the "rod movement" group dishabituated to the "broken rod" display while showing no dishabituation to the "complete rod" display. Thus, they generalized their habituation to the "complete rod" display while dishabituating to the "broken rod" display. On the other hand, the "baseline" group did not discriminate between the test stimuli.

The result was consistent with the interpretation that infants in the "rod movement group" interpreted the habituation stimulus as a complete rod moving behind the block. Any intrinsic preference, i.e. increase looking time, for either test display could be ruled out since the baseline group showed no preference. The study demonstrates that infants as young as 4 months already seem to know that the parts of the same object move together in a coherent fashion, i.e. that common motion indicates a single, connected object.

Another similar study Slater et al. (1990) indicates that newborn infants do not represent object unity. Naturally, the question arises how infants arrived at this skill between birth and four months. We will model this phenomenon and suggest our explanation in chapter 3.

1.2.4 Occlusion and object permanence

Intimately related to the problem of segmentation and object unity is the problem of occlusion. We may not be aware of this problem since "seeing things" comes to us so effortlessly. Research in computer vision has become painfully aware of these difficulties: when objects are occluded, what determines their edges? How do we know how the object looks like behind the occluder? How to deal with different lighting conditions? How to extract the true form of the object even though parts

of it are occluded? These questions pose severe problems in computer vision and they are currently far from being solved (see e.g. Azad et al. (2008)).

What interests us most is to understand how our effortless ability to extract (partly) occluded objects from a scene develops in infancy. Maybe then we will gain insight into adult occlusion perception. Interestingly, very young infants are not able to represent occluded objects or object parts - this so-called object permanence does not develop until the age of 2-6 months (Baillargeon et al., 1985, Baillargeon, 1987, Johnson et al., 2003). We will turn to the question of occlusion and object permanence in the chapters 2 and 3 and model the development of these phenomena in infancy.

1.2.5 Object categorization

Of course, the view that object unity perception can be fully accounted for by the common fate mechanism is highly simplified. How would an infant perceive unity of non-moving objects? These issues will be discussed later. As for now, let us turn to the question of how infants' cognition may develop after having acquired the ability to perceive objects.

Now that objects have been carved out from the chaotic pixel environment one might think of categorizing these objects. Intuitively one might think of grouping objects with similar properties together, like various books, chairs, houses, people etc. Classically, object categories have been defined at three different levels of abstraction: subordinate, basic and superordinate/global. For example, the subordinate-level category "Mercedes" encompassing all Mercedes cars belongs to the basic-level category "cars" which itself belongs to the superordinate-level category "vehicles" (Rosch et al., 1976).

Whether young children begin with the formation of categories at higher or lower levels of abstraction is still the subject of controversial discussions. Also, different scientific communities developed different non-compatible opinions on this question. For example, the computational neuroscience and computer vision communities prefer the view that objects and their categories are constructed from

their specific features such that higher level categories are acquired through abstraction from lower level categories, see e.g. Rolls and Deco (2002). This view is suggested by the current hierarchical and feedforward conception of the visual system. According to that view, visual information spreads from the retina to the lateral geniculate nucleus (LGN) and then to the most occipital part of our brain, the primary visual cortex V1. From there this information is thought to spread to higher areas while branching into two visual streams. Important for our discussion here is that the retina is thought to process the visual scene on a pixel level, V1 neurons have been shown to react to edges, i.e. arrays of pixels, V2 to edge configurations and finally at some point the inferotemporal cortex (IT) to whole objects, see e.g. Zigmond et al. (2003). Thus, it is not surprising that object recognition and categorization is viewed as a result of an complex abstraction process running through the processing stages of the visual system. According to this view, subordinate categories are expected to develop first, basic level later and the global level at last.

The history of opinions in the field of developmental psychology has been somewhat different. For a long time, basic-level categories were considered to develop first (prior to higher-order categories) because exemplars of a given basic-level category look more similar to each other than exemplars of a given higher-level category (Rosch and Mervis, 1975) and because the first category words uttered by infants tend to belong to the basic level (Mervis, 1987), i.e. “dog” is uttered earlier than “poodle” or “animal”. Contradicting this developmental hypothesis, later studies on categorization in infancy suggest the existence of a global-to-basic level shift during the first years of life (Mandler and McDonough, 1993, 1998). According to this idea, infants’ initial categories may be rather broad in nature.

In one of these studies infants were given plastic replicas of objects to freely explore, see Fig. 1.2. Each session consisted of eight familiarization trials and two test trials, e.g. a dog, bird, cat, horse, dog, bird, cat, horse and then a fish and an airplane. “fish” is a new object from the familiar category while “airplane” is a new object from a new category. The rationale of the study was the following: if infants consistently examined the new object from the new category (“airplane”)

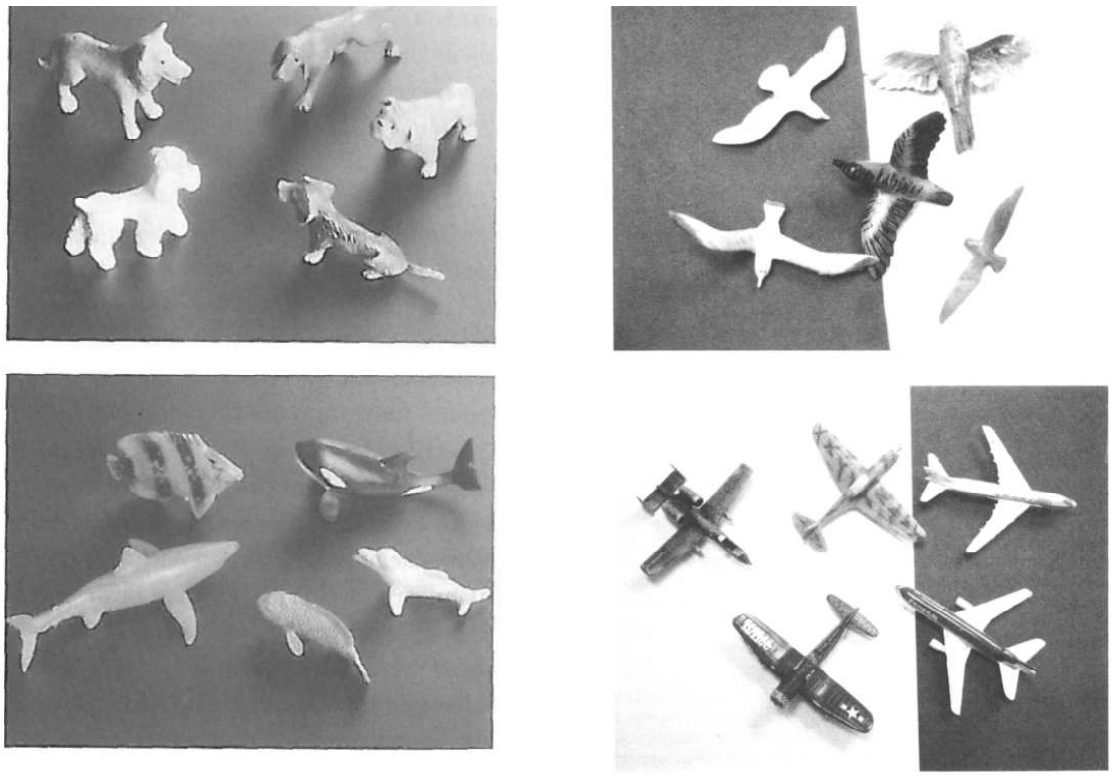

Figure 1.2: Objects for object-examining task in the categorization experiments (Mandler and McDonough, 1993).

significantly longer than the new object from the familiar category ("fish") then it can be concluded that infants are able to discriminate those two categories. Indeed, it turned out that 9-month-old infants are able to distinguish animals and vehicles, i.e. a distinction at the global level. This task is not easy to perform because solely on the basis of how the objects look like it is difficult to classify them into animals and vehicles. On the other hand, after performing a similar study at the basic level probing the ability to discriminate dogs from fishes, Fig. 1.2, left, infants failed. For example, infants would be familiarized with a series of eight dogs and then tested with a new dog and a fish. In a control study the researchers tested the discrimination of birds with spread out wings vs. airplanes. To the researchers' surprise, infants *did* discriminate them although they were perceptually very similar, see Fig. 1.2, right. Consistently Pauen (2002) showed that 8-month-olds are able to discriminate animals and furniture (global level) in an

object examination task while failing to discriminate dogs from birds and chairs from tables (basic level). Four months later, those infants, now 12 months old, succeeded in the discrimination at both levels.

In summary, infant research in the last two decades suggests that the direction of development in object categorization goes from global to basic to subordinate - the opposite direction as thought by the computational neuroscience and vision community. Specifically, this research shows that there is more going on in object categorization in infancy than just similarity judgments and abstractions. Moreover, it highlights that it may not be that important to recognize the details of objects in order to categorize them in meaningful entities.

Thinking more profoundly about this issue leads us to the following questions. First, why do infants from these global categories animals vs. artifacts before being able to distinguish e.g. animals from each other? And second, how is it possible to make these distinctions given that both animals look different from artifacts but also animals look different from each other and artifacts look different from each other? The straightforward answer that comes to mind is that objects are categorized based on the role they play in the world. It is about their actions and behavior in the world and not so much about their looks. After all, a house is not a house because it has vertical walls and a roof and windows etc. but because people live in it. Houses (chairs, people, furniture, vehicles,...) can look so different but humans are still able to categorize them effortlessly because they understand the role that these objects play in the world. Mandler (1992) suggests how this problem may be approached. For example, the distinction between animals and artifacts may be based on a primitive understanding of what living beings are vs. non-living ones, e.g. living beings move in a self-propelled way while artifacts usually move only when caused to.

This discussion leads us to the following research direction. It may not be important to recognize, categorize or understand details of the visual scene. This is even more true given that infants' visual resolution is very bad in the first months of life, see e.g. Kellman and Arterberry (1998). Maybe infants start, after segmenting the scene into objects, by putting them into global and meaningful categories

according to the role that the objects play in the world. The distinction of animate vs. inanimate or caused motion may be a starting point.

1.3 Outline of the thesis

Given the previous discussion, as a first step, we will attempt to model causality and occlusion perception in infancy in chapter 2 motivated by the conjecture that it reveals something about the first categorizations of animate vs. inanimate. We will try to build unified models that capture several phenomena because that in this way we may reveal important principles of cognitive development in infancy.

In Chapter 3 we will turn to the important problem of object unity which will turn out to be necessary to solve in order to proceed. Again, we will strive for unified models that also capture occlusion and object permanence and in principle even causality as well.

In Chapter 4 we will develop our thoughts further that lead us to the question of how infants learn regularities in the environment in the first place. This will be examined in Chapter 5 where we investigate the formation of visual expectations in sequence learning in infancy.

Finally, in Chapter 6 we will draw conclusions from this PhD work and outline areas of possible future progress.

Chapter 2

Development of causality and occlusion perception

2.1 Introduction

In this chapter we investigate the development of causality and occlusion perception in infancy. As for causality, the discussion can be traced back to Hume (1740) in whose classical account the perception of causality in simple mechanical events is the result of repeated experiences of a constant conjunction between two events. Michotte (1963) argued that causality could be perceived directly, for example, when one billiard ball collides with and launches another. He believed that in order to gain a "causal percept" infants would at least have to see enough "internal structure" to segregate a launching sequence into two movement components. Leslie (1994) believed that an innate notion of force or pressure is needed and that the perception of cause and effect is performed by an innate motion analysis module. Mandler (2004) suggested that seeing transfer of motion may provide the basis of infants' early interpretation of causal physical events and that no notions of force or pressure are necessary. We demonstrate that this is possible by constructing a computational model that learns to represent launching and occlusion events by merely observing them and detecting statistical regularities in them. We show that this model explains one of the fundamental experiments on the perception of causality in infants (Leslie, 1982) while no innate force notions or modules are needed.

The model is an artificial neural network that is trained to predict its future inputs. We are going to model two experiments on causality and occlusion perception that rely on the habituation paradigm (see Sect. 1.2.3). The network's error in predicting its next input is used to model the infant looking time since novelty can be seen as a prediction failure. Prediction learning has been highlighted by a number of developmental theorists (Elman, 1990, McClelland, 1995, Schlesinger and Young, 2003). To our knowledge, the first attempt to analogize the error in neural networks to infant looking time has been made by (Mareschal et al., 2000). We extend this line of work by additionally pre-training the network which is essential in this kind of modeling as will be argued in Sect. 2.2.2.

Using a connectionist prediction framework and we show that much of the basic physical knowledge need not be innate but can be learned with a small set of prior capacities.

2.2 Methods

2.2.1 General architecture

We used a simple recurrent network, also known as the Elman network (Elman, 1990). It consists of four layers of artificial neurons, named input, hidden, output and context layers, respectively (Fig. 2.1). The inputs to units in the hidden and output layers were weighted sums of the responses, X_i, C'_j, Y_j, from units at previous layers. The outputs Y_j and Z_i of the units were Fermi functions of the input:

$$Y_j = \left[1 + \exp\left(-\sum_{i=0}^{M} v_{ij} X_i + \sum_{j'=1}^{N} u_{j'j} C_{j'}\right)\right]^{-1} \tag{2.1}$$

and

$$Z_i = \left[1 + \exp\left(-\sum_{j=0}^{N} w_{ji} Y_j\right)\right]^{-1}, \tag{2.2}$$

where v_{ij}, $u_{j'j}$ and w_{ji} are the weights. Every hidden and output unit had an additional constant input X_0 and Y_0 equal to 1. The weights v_{0j} and w_{0i} of these

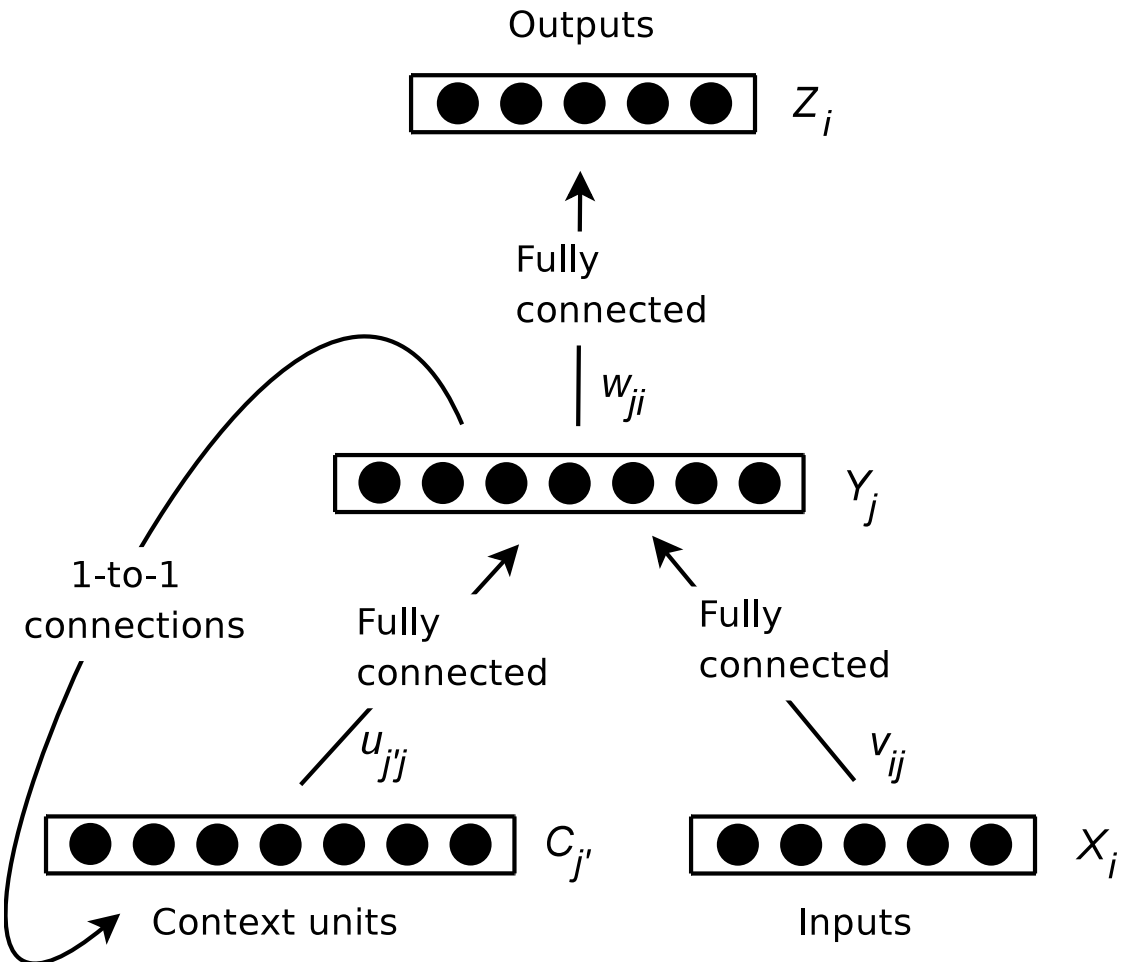

Figure 2.1: The Elman network

supplementary inputs act as threshold values for each unit and are also learned. The context layer derives its activity from the hidden layer by copying its activity at each time step: $C_j := Y_j$. The Elman network is presented a temporal series of inputs $X_i(t), X_i(t+1), X_i(t+2), ...$ and its task is to learn from this sequence and predict the next input $X_i(t+1)$. We trained the network with the standard backpropagation algorithm minimizing the sum of the squares of the difference between the output $Z_i(t)$ and the next input $X_i(t+1)$. In our model we relate the prediction error

$$E(t) \equiv \sum_{i=1}^{M} |Z_i(t-1) - X_i(t)| \tag{2.3}$$

to the looking time in experiments with infants[1].

Our model is constructed to predict occlusion and launching events. Therefore, it is necessary to represent motion and depth. In order to do so we split up the input layer into three maps, the "motion detectors" (first 7 units), "disparity" units (next 7 units) and "novelty" units (next 14 units) that represent the novelty of the environment. The network is used to model the visual events in Fig. 2.3. Fig. 2.2a) shows the inputs to the network at a launching event. The motion detector map is only active for a moving object (pixel). A unit is set to 1 if motion is present (in any direction) and to 0 otherwise. In order to predict occlusion events successfully it is necessary to distinguish at least three depth relations: farther away, same distance and closer than the object participating in the occlusion event. Therefore, units in the depth map can have three values, 0.0, 0.5 and 1.0, respectively. In Fig. 2.2g) we see an example of an occlusion event. The idea of the novelty map is that everything is new to an infant when it comes to the laboratory which leads to large looking times at first trials (see Sect. 2.2.2 for details).

2.2.2 Training

The network was trained in three phases: pre-training, habituation and test phase. In habituation studies infants are habituated to a repeating stimulus until the looking time drops and the habituation phase is terminated. Then infants are presented test stimuli. Our model was trained in a similar way except that a pre-training phase is needed. The pre-training models the visual experience of the infant since it had experience with the world before coming to the laboratory. Without pre-training the network would only learn what is presented during the habituation phase, from which no interesting results can be expected.

We pre-trained the network in the following way. A freely moving (not occluded) pixel moved back and forth as displayed in Fig. 2.2h). This motion was halted and reinitiated with probability of 5% at each time step. Another non-moving pixel was added or removed with probability of 1% (at each time step) at a random

[1]The prediction error could have as well been defined as the objective function (sum of *squared* errors) which is simpler to treat mathematically because of the well-defined derivative.

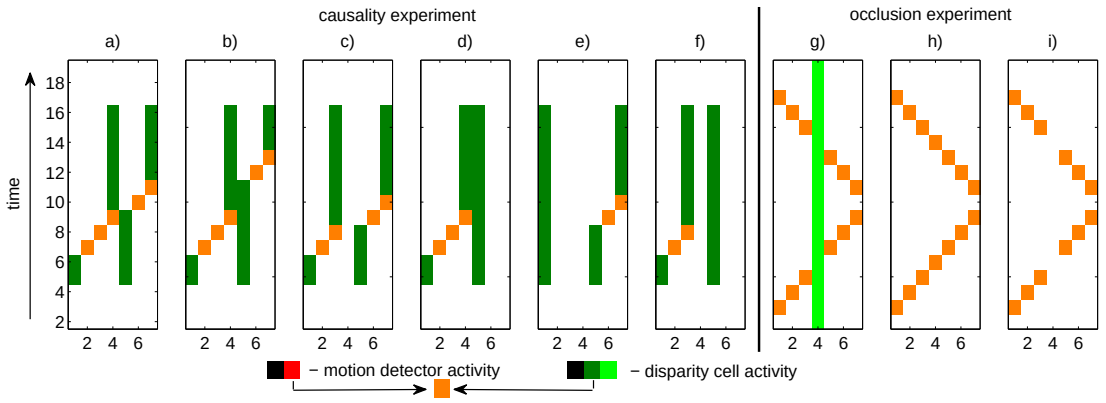

Figure 2.2: Habituation (a,b,c,g) and test (d,e,f,h,i) stimuli. Motion detector and disparity layers are displayed on top of each other, red pixels indicating motion detector activity and green scale values indicating disparity cell activity. Note that whenever a motion detector is active the corresponding disparity cell's activity is 0.5 (resulting in the mixed color orange). a) *direct launching* - first pixel launches the second pixel. b) *delayed launching* - second pixel moves off 2 time steps after collision. c) *launching-without-collision* - second pixel moves off without being touched. d) *no-reaction* - first pixel collides with second pixel which fails to move. e) *no-prior-movement* - second pixel moves without prior movement of first pixel. f) *no-reaction-no-collision* - first pixel stops before touching the second pixel which remains inert. g) *occluded trajectory* - pixel moves back and forth behind an occluder. h) *continuous trajectory* - pixel moves freely without occlusions or collisions. i) *discontinuous trajectory* - interrupted pixel motion

position. When this second pixel was there, its depth was chosen to be 0.5 or 1.0 with equal probability, so that the pixel either became an obstacle or an occluder. In the occluder case the stimuli were like in Fig. 2.2g) where occlusions could occur at any position. In the obstacle case a launching event occured when the first (moving) pixel collided with the second one (see Fig. 2.2a)). During pre-training the occluder or obstacle could be at any position in the visual field, while during the habituation and test phases the stimuli were exactly those shown in Fig. 2.2. The stimuli during the pre-training were constrained to "possible" ones, i.e. to stimuli that we would expect to occur naturally like direct launching, occlusion or just free motion. By doing this we model the infant's pre-experimental experience with the world. Varying the pre-training time allows us to look inside the development of the model - therefore, the pre-training time corresponds to the age of the infant. The novelty units are set to a constant but random binary vector. The

network was pre-trained for 5×10^5 time steps and the weights were saved every 1000 time steps. Then we performed the experiments described below using these saved weights that represent the developmental progress of the network.

2.3 Perception of causality: modeling Experiment 1 by Leslie (1982)

2.3.1 Description of the original experiment

Leslie (1982) tested how 4.5- and 8-month-old infants perceive a launching event. The stimuli are shown in Fig. 2.3a)-f). Infants were habituated to a cube starting to move and launching another cube which starts to move with the same speed as the first one while the first one stops moving after the collision (direct launching). Another group of infants was habituated to the same stimuli except that the start of motion of the second cube was delayed (delayed launching). A third group of infants was presented a launching event without the first cube touching the second one. It stopped at some distance before but the second cube started to move off immediately just as if it had been launched (launching without collision). All infants were tested with basically two kinds of events: first cube moving, colliding with the second, stopping but without any reaction of the second cube (no-reaction). Alternatively, the second cube just started to move by itself without prior motion of the first cube (no-prior-movement). In Fig. 2.2a) - f) we see how these events have been presented to the neural network.

Leslie investigated whether infants perceive the first cube as causing the second cube to move. The idea was that introducing a temporal or spatial gap between the two cubes would make the infants perceive two independent motions: one cube staring, moving and stopping and then the second one doing the same - which corresponds to the test stimuli. Therefore, Leslie hypothesized that infants who were habituated with the delayed launching or launching-without-collision sequence would dishabituate less to the test stimuli than the infants exposed to the direct launching sequence. We will discuss the results together with the modeling

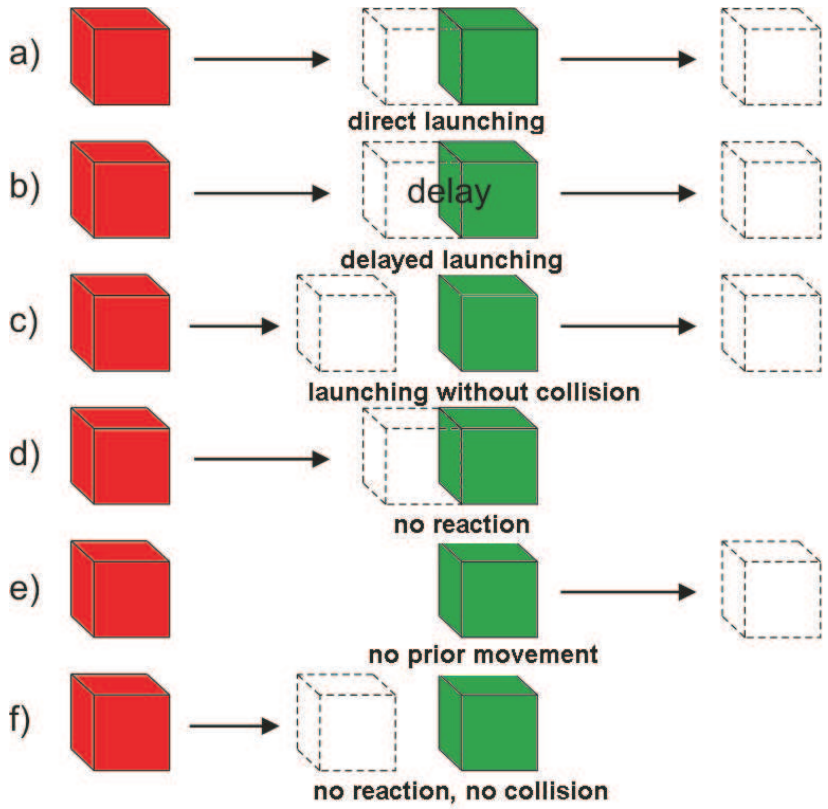

Figure 2.3: Stimuli in Experiment 1 from Leslie (1982).

results below in Sect. 2.3.3.

2.3.2 Modeling procedure

After the pre-training the network was trained repeatedly with the direct launching stimulus in Fig. 2.2a) (alternately b) or c)) for 100 time steps (habituation phase). Then, the total prediction error which is the sum of the prediction error over the 18 time steps of a stimulus (see Fig. 2.2) was calculated. After habituation, the test stimuli were presented once and the prediction error was calculated again. The direct and delayed launching habituations were tested with the no-reaction and no-prior-movement stimuli whereas the launching-without-collision habituation was tested with the no-reaction-no-collision and the no-prior-movement stimuli thereby modeling the original study procedure. During the habituation and test trials the novelty units were switched to a different binary random but con-

stant vector indicating the novelty of the laboratory environment (the novelty units don't play any role in this experiment but are important for the control case in the occlusion experiment below).

2.3.3 Results

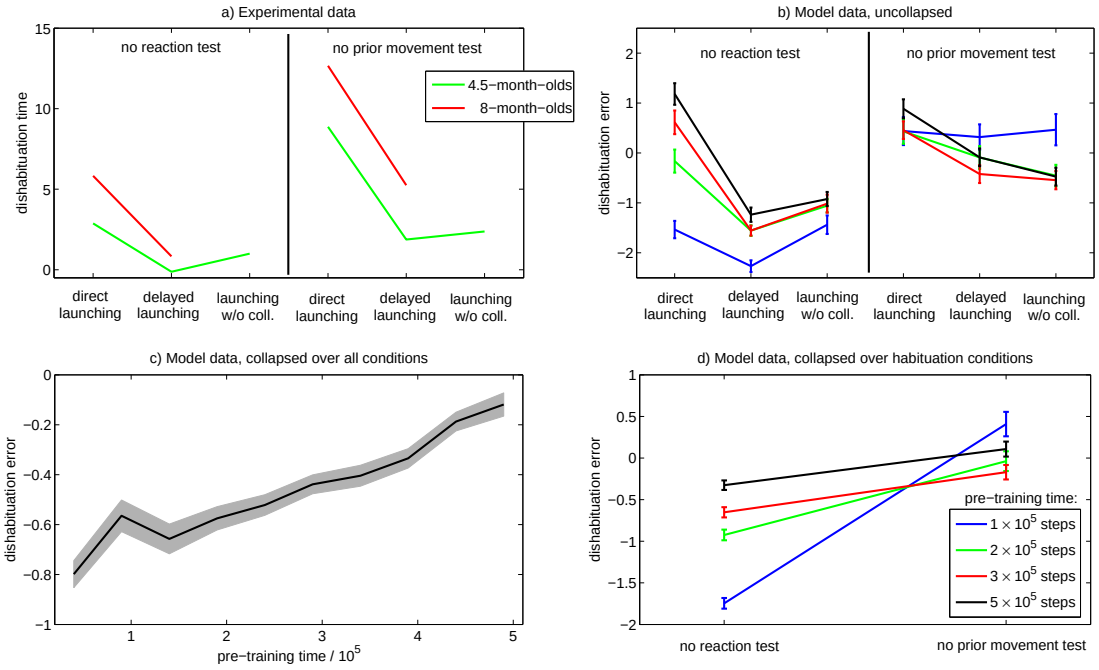

Figure 2.4: a) Dishabituation times in the causality study (Leslie, 1982). For analysis of statistical significance see original publication. b) Dishabituation errors in the model. c) Dishabituation errors as a function of pre-training time. d) Dishabituation errors collapsed over habituation conditions.

The whole simulation was run 40 times. In Fig. 2.4 the results together with the results of the original experiment are shown.

Experimental result:

Looking times declined significantly from first to last habituation trials.

Model account:

During the habituation phase the stimulus was repeated over and over (100 / 18

$\approx$ 5-6 trials). Thus, the network learned to predict the stimulus better, i.e. its prediction error dropped with time. As we relate the prediction error to the looking time of infants this accounts for this result.

Experimental result:

The group shown the direct launching stimulus increased its looking time significantly more than the group shown delayed launching or launching-without-collision.

Model account:

Since the network was exposed to direct launching stimuli during pre-training already, there was not much to learn during the direct launching habituation. On the contrary, the other two habituation stimuli were more difficult to learn. Thus, the prediction errors of the last habituation trials were lowest in the direct launching case. Therefore, the network's "dishabituation" to the no-reaction test was higher after direct launching as compared to the other habituation cases (see Fig. 2.4b).

Experimental result:

The no-prior-movement stimulus attracted significantly longer looking times than the no-reaction stimulus, regardless of the group.

Model account:

During habituation the network learned to expect the first pixel to start moving which happens in the no-reaction test but does not happen in the no-prior-movement test. Thus, in the no-prior-movement test, the network keeps predicting the first pixel to start moving which does not happen and yields an additional prediction error (see Fig. 2.4d).

2.3.4 Model predictions

Although Leslie did not find any significant age effects, we see in his results, Fig. 2.4a), that the mean dishabituation time is higher for older infants (which could be random, of course). On the contrary, as displayed in Fig. 2.4c), our model predicts that overall dishabituation times increase with age which is due to the fact that the habituation stimuli can be learned quicker and better after a long pre-training. Another prediction is that the direct launching results should be more similar to the delayed launching and launching-without-collision results

for younger infants. This is due to the fact that the dishabituation errors are higher in the direct launching condition only because of prior exposure to direct launching stimuli during pre-training. If, on the other hand, pre-training is short (young infants) then this effect vanishes as confirmed by Fig. 2.4b).

2.4 Perception of occlusion: modeling Experiment 1 by Johnson et al. (2003)

2.4.1 Description of the original experiment

The experimenters tested 4- and 6-month-old infants and discovered an interesting effect in occlusion perception. They habituated the infants with a ball oscillating back and forth behind an occluder. Then two kinds of test displays were presented, both without any occluder: the first test display showed the ball continuously oscillating and the second test display showed the ball oscillating discontinuously, i.e. they used the same display as in the habituation but they removed the occluder such that it appeared that the ball oscillates but disappears behind an invisible occluder and reappears at its other end again. A separate control group was shown the same test stimuli but without prior habituation in order to control for some baseline preference for one of the test displays. In Fig. 2.2 we see the corresponding habituation, g), and test stimuli, h) and i), that we used in our model.

If infants perceive the ball as continuing to move behind the occluder during habituation then they should generalize their habituation to the continuously moving ball and show increased looking time at the discontinuous test display. However, if infants just learn the motion of the ball "by heart" then they should generalize this perception to the discontinuous case which is identical to the ball's motion in the habituation display. Thus, they should dishabituate more to the continuous display. The experimenters found that 4-month-olds show a preference for the continuous display but 6-month-olds dishabituate more to the discontinuous display (Fig. 2.5a)).

2.4.2 Modeling procedure

After pre-training the network was habituated repeatedly with the stimulus in Fig. 2.2g) for 1000 time steps. After habituation the network was tested again once with the two stimuli Fig. 2.2h) and i) and the prediction errors, $E^{\text{exp}}_{\text{cont}}$ and $E^{\text{exp}}_{\text{discont}}$ were calculated. The prediction error was also calculated for the last habituation trial, $E^{\text{exp}}_{\text{baseline}}$, in order to be able to calculate the dishabituation time later.

Just as in the real experiment we also modeled the situation of the control group, i.e. we took the pre-trained weights, tested them directly without prior habituation and calculated again the prediction errors $E^{\text{control}}_{\text{cont}}$ and $E^{\text{control}}_{\text{discont}}$. In order to assess the dishabituation errors we presented the occlusion display 2.2g) once to the pre-trained network having still the old novelty units (see Sect. 2.4.3 for the role of the novelty units). This prediction error, $E^{\text{control}}_{\text{baseline}}$, reflects how the network had learned so far.

Finally, we calculated the "looking preferences", P, for the experimental and control conditions, respectively.

$$P \equiv \frac{E_{\text{discont}} - E_{\text{baseline}}}{(E_{\text{cont}} - E_{\text{baseline}}) + (E_{\text{discont}} - E_{\text{baseline}})} \tag{2.4}$$

The difference between the test error and the baseline error (last habituation error) is what we call the dishabituation error which is analogous to the dishabituation time in real experiments. In the original experiment the researchers used the same formula for the preference except the baseline values, i.e. they took the raw looking times to the test stimuli.

2.4.3 Results

The whole simulation was run 40 times. In Fig. 2.5 the model as well as the experimental results are shown.

Experimental result:
4-month-old show a preference for the continuous display whereas 6-month-olds

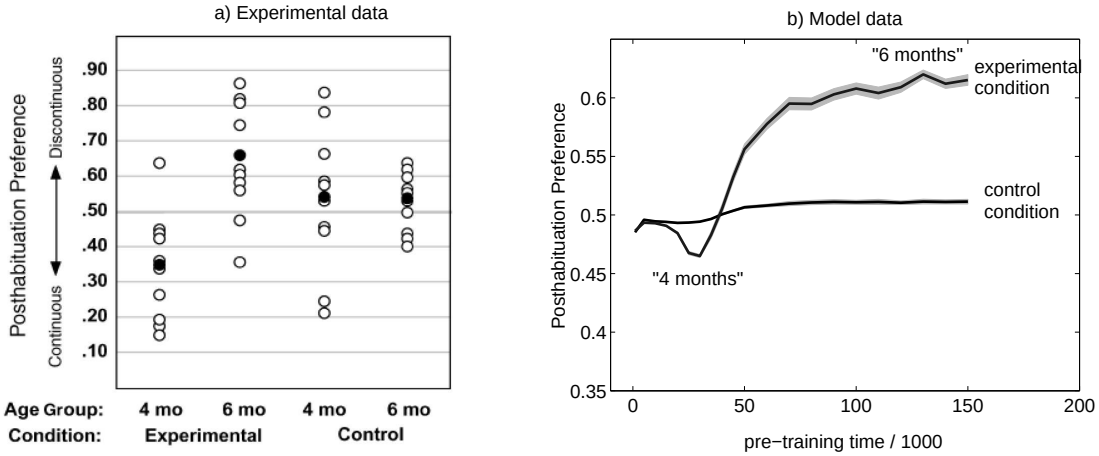

Figure 2.5: Development of looking preferences for the continuous vs. discontinuous displays in a) experiment and b) model. The gray shaded areas denote error bars based on 40 simulations.

prefer the discontinuous display.

Model account:

In Fig. 2.5 we see that the preference first goes down to about 0.46 after 30000 time steps and then increases until is saturates at 0.62 as a function of the pre-training time steps. Since the pre-training time corresponds to the infant age we get a similar result as observed experimentally. The preference curve can be explained in the following way: After 0 time steps the network did not predict any output at all. Therefore, the test displays are equal to the errors which are both large and lead to a preference around 0.5. After 30000 time steps the network has learned to predict the trajectory of the pixel except at the occlusion position to some extent. It basically learned the pixel motion "by heart". This is the same case as before but only with smaller errors in total such that the failure to predict the continuous trajectory at the occlusion position gained more weight (preference for the continuous display increased). After the network has been exposed long enough to pre-training stimuli it learned to predict continuous trajectories and also that a trajectory is suppressed whenever there is an occluder. Therefore the network predicts the continuous display successfully but fails to predict the discontinuous one since it expects the pixel to move on in the absence of an occluder (preference for the discontinuous display).

Experimental result:

The control group showed no preference for either test display.
Model account:
This is due to the general novelty of the laboratory environment. As the network was shown test stimuli without prior habituation it could not learn the value of the new novelty units which increases the total prediction error. Of course the network did have a baseline preference for the discontinuous ("unnatural") display after a long pre-training time. But this difference was small as compared to the prediction errors of the novelty units and reduced the total preference to around 0.5. Therefore we suspect that there may be a baseline preference also in infants but the novelty of the laboratory environment makes any preference non-significant.

Object permanence
In contrary to real experiments with infants we can examine *how* the system achieves its performance. First it learns to predict freely moving pixels in either direction. But then, when an occluder is present, it simply learned to suppress the output of the motion detector layer at the position of the occluder. This is done by a few hidden units that are mainly driven by the activity of the disparity layer cell that represents the existence of an occluder. Whenever these hidden units are active they suppress the activity of the motion detector output layer at the same position where the occluder occurred by feeding in strongly negative connection weights to them. In this way the other hidden layer units *do* predict a moving pixel to continue its motion at the occluder position but the actual output of this prediction is suppressed by the former hidden units. Therefore, the network continues to represent the motion of the pixel even though there is an occluder which is exactly what object permanence means. We did not foresee this capacity of the network to develop, it just emerged as a solution to the occlusion problem.

2.4.4 Model predictions

As we can see from Fig. 2.5b) the model predicts that infants' preference should be similar to the model curve if the experiment will be performed for other ages as well.

2.5 Discussion

In the original causality experiment Leslie (1982) asked whether infants perceive two cube movements, one continuously causing to move the other. They found that the perception of "causality" is disrupted if a temporal or a spatial gap is introduced between cause and effect. A gap would supposedly lead to the perception of two separate, independent movements whereas a direct launching is supposed to be perceived as two conjoined movements.

The model can be said to provide a similar but a much more precise account. There are no two separate movements in the model. There are just sequences of input vectors. But the model learned that whenever a pixel approaches and touches a resting pixel it stops and makes the latter move in the same direction. Specifically, it learned that the second one moves off immediately (direct launching). Therefore, the model "dishabituates" more when presented a no-reaction test as compared to the condition where it has been habituated to delayed launching. Thus, there is no need for innate force notions of motion analysis modules as has been proposed by Leslie (1994) whose experiment we modeled. Our model accounts for his data while suggesting that causality can be learned by merely observing the visual environment, registering statistical regularities and trying to predict them.

In the occlusion experiment (Johnson et al., 2003) the researchers wanted to know whether infants are able to perceive a continuous trajectory although partially occluded. They found that 6-month-olds seem to perceive the trajectory veridically but 4-month-olds do not - they rather seem to perceive two distinct sections of the trajectory. This makes sense with regard to our model. Before learning enough about the continuity of a trajectory the model/infant can not know that it must be continuing behind an occluder. Only after being able to successfully predict a free trajectory the road is free to follow an object behind an occluder with the "mind's eye". This is exactly what happened in the model and object permanence emerged as discussed above.

In summary, we presented a simple framework - a network trying to predict its

future inputs - that was able to develop representations for causality and occlusion perception as well as object permanence. It has learned about the continuity of object motion, about solidity and reaction of objects to contact. Therefore, there is no need to postulate innate principles as had been suggested by Spelke (1994). In our model we show that all these properties can be simply derived from statistical motion properties of raw visual input.

One drawback of the model is that it uses backpropagation of an error signal which is not biologically plausible if we want the framework to be a model of the infant brain. This is certainly a weakness but can be overcome by just using a more plausible network, e.g. a dynamic reservoir network with spiking neurons that have been shown to be able to perform prediction tasks (Lazar et al., 2007).

A major topic of future work will be to provide the model with an active representation of occluded scenes. Although this model shows object permanence, i.e. it represents occluded inputs, it cannot do so for more than one time step and it cannot habituate to what it merely represents and can only habituate to actual, real inputs. This is important since many experiments rely on the observation that infants apparently can habituate to things they don't see but only represent. This issue will be addressed in the next chapter as we model object unity and occlusion perception where we continue our effort to unify different aspects infants' basic physical knowledge about the world in a single computational theory.

Chapter 3

Development of object unity, object permanence and occlusion perception

3.1 Introduction

In the previous chapter we investigated how infants may perceive launching and occlusion events. In doing so, we abstracted from the visual appearance of the involved objects to single pixels in the input layer of our artificial neural network. This begs the question of how infants arrive at this level of abstraction in the first place, i.e. it is important to investigate how infants arrive at perceiving objects at all.

Therefore, understanding that our world consists of objects that move coherently on continuous paths and that keep existing even when occluded is an important step in our development. For example, research on the perception of object unity seems to have converged on the result that by roughly four months infants show the ability to interpret a rod moving behind a block as complete although only the rod ends are visible (Kellman and Spelke, 1983, Johnson and Aslin, 1995, Johnson and Náñez, 1995), see Fig. 1.1. Some developmental researchers concluded that the perception of object unity is one of the pieces of "core knowledge" that infants are hypothesized to be innately endowed with (Spelke, 1990). On the other hand,

there is evidence that newborn infants are not able to perceive object unity (Slater et al., 1990), which makes it difficult to explain the phenomenon in nativist terms.

Another ability that develops early in infancy is object permanence (Baillargeon et al., 1985, Baillargeon, 1987) and tracking of occluded object trajectories (Johnson et al., 2003, von Hofsten et al., 2007) that we started to investigate in the previous chapter.

Most prior studies and theoretical models highlighted only one of these aspects of an infant's perception or cognition. A unified model of object unity and occlusion perception has not been suggested yet, which is important since it can potentially reveal general principles that guide development. In this chapter we attempt to make another step in this direction and to provide a computational model of the development of object unity, object permanence, and tracking of objects behind occluders. We present an artificial recurrent neural network that is trained with different occlusion events, learns to represent occluded parts of objects, "perceiving" their unity, can keep the representation over short time intervals and also track objects that disappear and reappear behind occluders. In order to model infants' learning from the environment, we pre-train the network with stimuli that might be analogous to the ones that infants are exposed to during the development. Importantly, the network is not pre-trained differently for each experiment since infants are not prepared for the specific studies either.[1] Conversely, we pre-train the network only once and then perform all experiments. After pre-training we expose the network to habituation and test stimuli and compare its performance with infant data. The principle that guides the learning of the network is the prediction of future inputs (Elman, 1990). We show that this approach can successfully explain in total 12 studies covering several aspects of infants' visual development and is also able to make specific predictions for future studies.

In this chapter our model replicates and extends the ideas of the model in the previous chapter on object tracking behind occluders. Spefically, the development of object tracking behind occuders of varying sizes will be investigated here. We

[1] Use of language: throughout this chapter the term "experiment" is meant to refer to modeling experiments as opposed to the term "study" which denotes infant experiments.

refrain from modeling causality perception here although it could still be embedded in the present framework (see Sect. 3.7.3 for discussion).

Assumptions

In order for the model to be most helpful for the developmental researcher, we summarize here our assumptions.

1. **Infants acquire the abilities in all three domains (object unity, object permanence and occlusion) by learning spatio-temporal correlations in the environment, whereas learning is guided by the same principle: the prediction of future inputs / events.** The way this can be achieved is demonstrated by the model.

2. **During development infants are exposed to both occlusions and unoccluded moving and stationary objects. Objects occur in all combinations, i.e. moving or stationary objects occur in front of and behind other moving or stationary objects.** This assumption should be obvious since occlusions of moving or stationary stimuli are indeed ubiquitous in everyday visual environment.

3. **During the habituation studies, infants form interpretations of the habituation stimulus which are compared to (interpretations of) the test stimulus. The difference between them drives their looking time.** This is a broadly accepted way of interpreting habituation experiments with occluded stimuli, see e.g. Baillargeon (1999), Johnson et al. (2003), Kellman and Spelke (1983).

4. **Stimuli with different motion direction as well as background and foreground stimuli are processed by different neural populations.** The choice of neurons tuned to motion direction is biologically plausible since there are velocity tuned cells in the primary visual cortex (Orban et al., 1986). Moreover, it is justified to assume the model's sensitivity to motion since it is well established that neonates react strongly to moving stimuli (Slater et al.,

1985). As for the background-foreground separation there is some evidence that neonates are already able to do simple figure-ground segregation (Slater et al., 1990). Although this does not necessarily entail that background and foreground stimuli are generally processed by different neural populations, it is still true for most cases where foreground and background stimuli are in motion relatively to each other and are therefore processed by different velocity-tuned neurons as mentioned above. Moreover, the separation of the hidden layer into a foreground and a background layer is not strictly necessary and has only technical reasons since learning in such a large network is difficult. In the previous chapter we already demonstrated that object permanence and tracking can be modeled with a single hidden layer.

In order to test whether this set of assumptions is indeed sufficient to explain the development of object unity, permanence and tracking, we instantiated the assumptions in a simple recurrent network model.

3.2 Computational model

3.2.1 General architecture

The computational model[2] consists of two simple recurrent networks that are trained to predict their next inputs (Elman, 1990) as depicted in Fig. 3.1. The input layers are presented binary vectors of activity which are fed to the hidden and output layers consisting of sigmoidal neurons (for details see Sect. 3.8.1). At each time step, the hidden layer activities are copied into the respective context layers which feed the activity back to the hidden layers via trainable weights. This is the only place where recurrency is introduced. Each visual scene is divided into six 7×7 pixel maps, three background maps and three foreground maps, see Fig. 3.2. The background and foreground maps consist of left motion, right motion and stationary maps, respectively, resulting in $7 \times 7 \times 6 = 294$ input units. Each map can be thought of consisting of velocity tuned neurons, i.e. if a partly occluded rod moves to the right then the "right motion" background map and the "stationary"

[2]MATLAB code is available upon request.

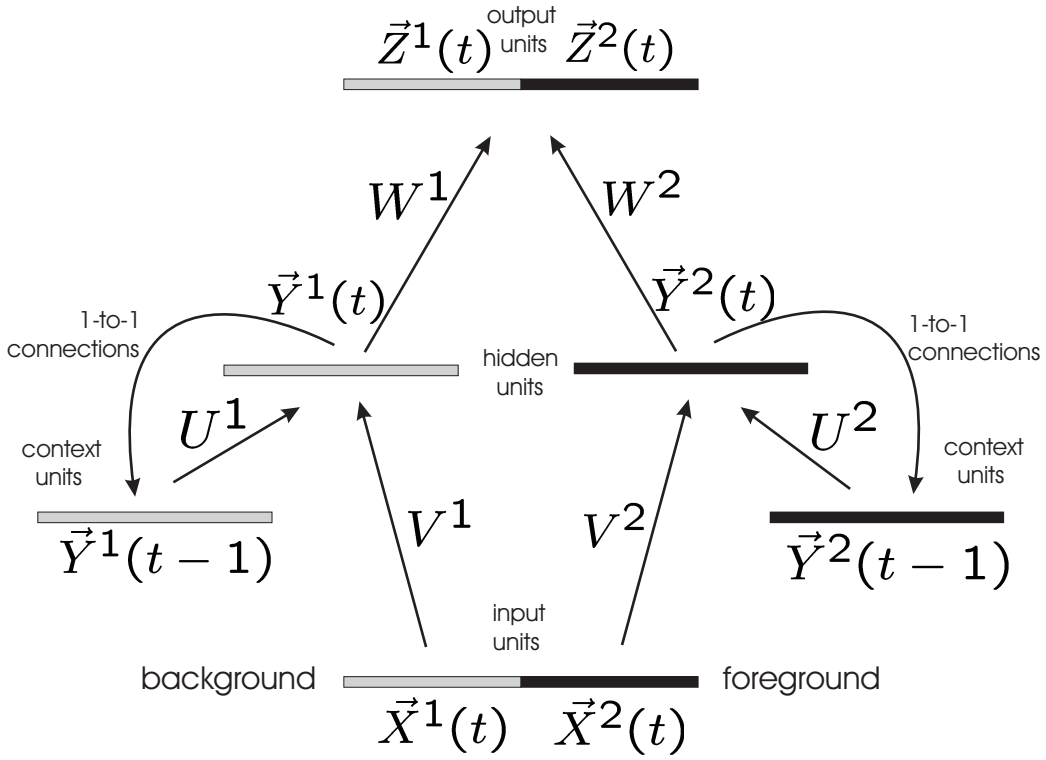

Figure 3.1: The network architecture consists of two simple recurrent networks (gray and black) with a common output layer. The network is trained by back-propagation through time to predict future inputs.

foreground map encode the stimulus.

3.2.2 Pre-training

Before coming to the laboratory for an experimental study, infants have gathered plenty of visual experience. Therefore, we include a pre-training phase where the network is presented stimuli that model the infants' visual experience before coming to the laboratory. These stimuli are unrelated to the experiments. The pre-training phase is necessary for modeling habituation studies during development since the behavior during the studies has to depend on prior knowledge.

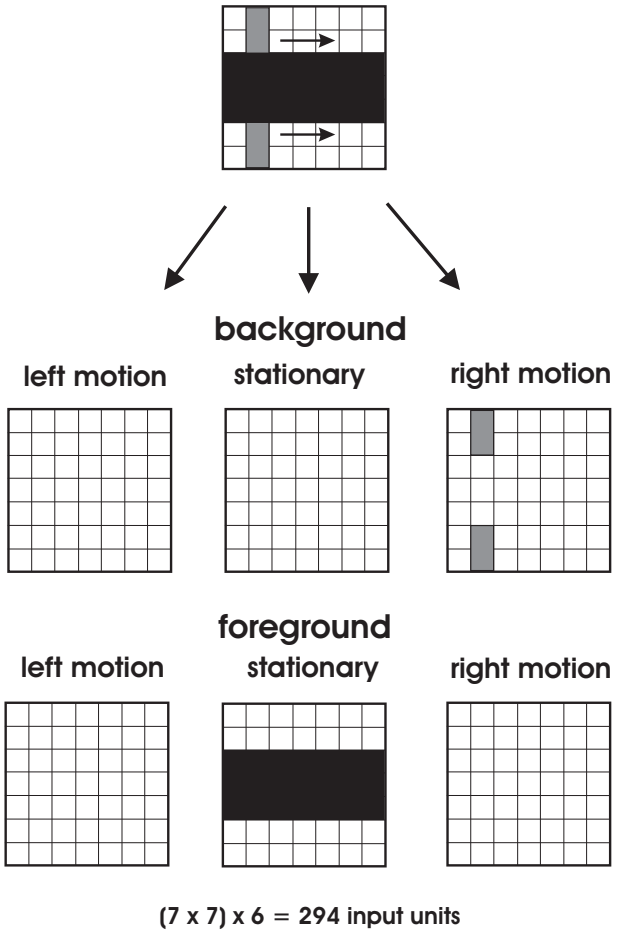

Figure 3.2: Input coding in the network via six maps where each map is a 7 × 7 dimensional binary vector. Each stimulus is segregated into foreground and background, each consisting of "left motion", "right motion" and "stationary" maps.

The network is trained to predict the input in the next time step. For training, the backpropagation through time algorithm is used (Hertz et al., 1991), see Sect. 3.8.2 for details.

During pre-training the network is presented rectangular objects moving and occluding each other or moving freely without occluder, see e.g. Fig. 3.3, "rod movement" or "complete rod" displays. The underlying reasoning is that once the network learns to predict a freely moving (not occluded) object, it keeps its rep-

resentation in the hidden layer even though the object disappears during occlusion. Note that background and foreground are encoded by distinct hidden layers. Therefore, during occlusion the background hidden layer has no information about the occluder, its input disappears for the time of occlusion and reappears at the opposite edge of the occluder. As we will see, the hidden layer learns to keep its representation of the moving (background) object until it reappears again. Note that the foreground and background parts of the network are completely symmetric and are also treated in a symmetric way.

The inputs during pre-training are constructed in the following way. The network inputs are initialized to displaying two objects, one in the background and one in the foreground. Each object is switched on with the probability of 0.7, thus, sometimes both objects are present, sometimes only one and sometimes no object. In the latter case there is no input to the network. Every object is a rectangle with a specific width, height, velocity (moving one pixel per time step to the left, to the right or stationary) and position in the 7×7 pixel input grid. From the two objects one is labeled "narrow" and the other one as "wide". A narrow object has always width one, while its height h is drawn from the uniform distribution $\{1, 2, ..., 7\}$. As for the wide object, the width is drawn from the uniform distribution $\{1, 2, ..., 7\}$ and the height similarly from $\{1, 2, ..., 5\}$. The y-position of every object is chosen such that the object is centered in the input grid, i.e. $y \equiv 4 - \text{floor}(h/2)$. Also the velocity of each object is picked from the uniform distribution on $\{-1, 0, +1\}$.

Having configured the objects and their properties the pre-training is started. At each time step this configuration is reset with probability of 0.04 according to the constraints defined above, i.e. most of the time the configured objects stay and move the way they are defined to, but once in a while the whole configuration is changed. The change involves a new pick of the objects' properties (width, height, velocity and position) and swapping the labels "narrow" and "wide" which put the narrow object sometimes in the foreground and sometimes in the background. For moving objects the x-position performs an oscillation around the central position 4 with amplitude 7, i.e. every time x reaches the position -3 or 11, the sign of the velocity is reversed. Note that this amplitude is large enough for the objects to

disappear from the display of width 7.

3.2.3 Habituation

Comparable to infants, the network is exposed to habituation and test phases after pre-training, see e.g. Fig. 3.3, left. During habituation a specific stimulus is shown for 1000 time steps while learning is still switched on. This corresponds to 62 trials roughly, since each trial consists of 16 time steps. This is apparently not in a realistic range but it reflects a general difficulty of learning in neural networks: the relation of the number of weight updates to real time is arbitrary. Note that without pre-training the network would only learn from the habituation stimulus from which no interesting results can be expected.

3.2.4 Tests and performance measurement

After habituation two particular test sequences are presented to the network. During habituation it is conjectured that infants interpret the visual scene in a particular way. This interpretation must be kept in short term memory until the tests are presented. If the infant generalizes from the habituation stimulus to the test stimulus, it will not dishabituate to it, i.e. the looking time will not increase. Otherwise the infant is surprised, which results in an increase in looking time.

By analogy to infants, we construct the following performance measure in the computational model. Suppose a rod moves behind a block as depicted in Fig. 3.3, "rod movement" display. If the network learns well it develops a representation of the complete rod in the background hidden layer even though the central part is missing. Thus, it infers and internally represents a complete rod during habituation. To assess the performance, the difference between this representation and the test displays is computed (see Sect. 3.8.3 for details). If e.g. the representation corresponds to a complete rod and the test display shows a complete rod as well then this difference will be small. On the other hand, if the test display shows a broken rod, the difference will be large. In the following, this difference will be called dishabituation error, E, in analogy to the dishabituation time in infant

studies.

For every test this yields two dishabituation errors, E_1 and E_2. The network's preference, P, is then given by

$$P = \frac{E_2}{E_1 + E_2}. \tag{3.1}$$

Eq. (3.1) is the same performance measure as used in some infant studies (Johnson et al., 2003).

3.3 Perception of object unity

The perception of object unity is an important aspect of the understanding of visual scenes. If we did not know which parts of the visual input belong to the same object we could not categorize the world into objects and assign properties to them. Because this capacity is so crucial many studies have been conducted on this topic. In this section we will present some of those studies, the results of our model and make predictions about new studies.

3.3.1 How does the perception of object unity develop?

S1: Original studies

In Sect. 1.2.3 we already discussed the original study on object unity by Kellman and Spelke (1983) with the main result being that 4-month-old infants seem to be able to perceive a united, complete rod as moving behind a block. On the other hand, a similar study with newborn infants demonstrated that infants do not possess this ability from birth (Slater et al., 1990). Neonates behaved in the *opposite* way as compared to 4-month-olds, i.e. they showed a preference for the "complete rod" display. The authors concluded that newborn infants do not perceive the rod pieces as connected but are nevertheless able to discriminate figure from ground, otherwise they would not have shown any preference.

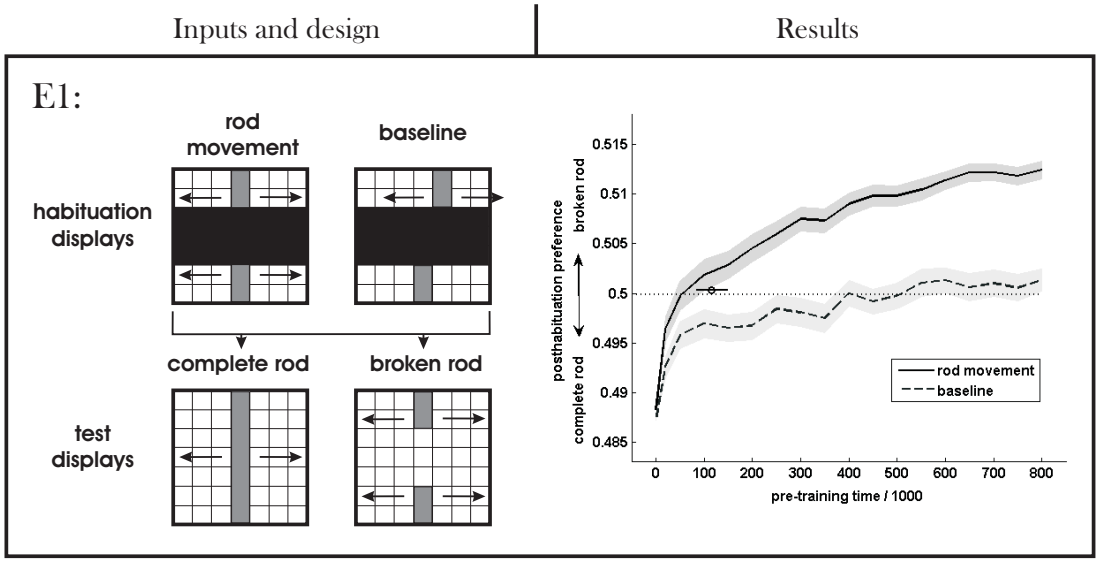

Figure 3.3: Left: input displays and design in modeling Experiment 1, compare Fig. 1.1. Background (gray), foreground (black). In the "rod movement" condition the rod starts on the left side of the display, moves to the right one pixel at a time, disappears for one time step at the right edge and then comes back performing the same motion to the left. The central part between the rod ends is never visible. Right: dishabituation preferences of the model in Experiment 1 as a function of pre-training time. The gray shaded areas denote the standard error bars calculated on the basis of 25 simulations. The circle indicates where the "rod movement" preference curve intersects 0.5 and the line through the circle denotes a standard error bar indicating to which precision this value is known. Note that this intersection is not where it seemingly should be according to the figure. See Sect. 3.8.4 for explanation.

Since neonates behave in the opposite way as compared to 4-month-olds there should be an age where infants undergo a transition period assuming that development is continuous. Indeed, Johnson and Náñez (1995) provide evidence for such a transition period: 2-month-olds were investigated with a similar rod & block design and did not show any preference for either test display. Given the evidence, it seems that the development of object unity perception is a continuous process that operates in the first 4 months of life.

Modeling procedure

We modeled these studies by first pre-training the network with different visual stimuli as described above. In this way infants' visual experience prior to the study is taken into account. By varying the length of the pre-training, infant behavior at different ages can be modeled. After pre-training, the network is trained with the habituation stimulus for 1000 time steps. By analogy to the original stimuli in Fig. 1.1 the network is trained with the stimuli in Fig. 3.3. To be precise, the network is trained with the "rod movement" stimulus and a copy of the network is trained with the "baseline" stimulus. This corresponds to different infant groups in the study. After habituation both test stimuli are presented to each copy of the network and their preferences, P, are calculated as described in eq. (3.1).

E1: Results

Fig. 3.3 shows the network's preference as a function of pre-training time. The network that is habituated to the "rod movement" display begins with a preference for the "complete rod" display, undergoes a transition period at 115000 time steps and finally develops a clear preference for the "broken rod" display while the "baseline" network saturates at 0.5 roughly — no preference for either test display. This is in agreement with the picture derived from the studies above (it should be mentioned that the baseline condition in Fig. 1.1 was never actually tested by Slater et al. (1990)).

Unaligned edges

There is also data on the case if the rod ends in Fig. 1.1, "rod movement" display, are not aligned (Johnson and Aslin, 1996). In that case 4-month-old infants preferred the complete rod (in the nonrelatable case, i.e. edges could not be extended to meet behind the occluder), i.e. infants seem to have interpreted the display as disconnected. This data is captured by the model as well since rod ends that are not aligned never occur during pre-training and therefore no correlated "completion" can be constructed by the network. This issue is beyond the scope of this

thesis though.

3.3.2 Unity perception for stationary objects

S2: Original study

Looking at the "rod movement" condition in Fig. 1.1 instantly creates the impression that even if the rod pieces don't move there is a complete rod behind a block. Why is that the case? It seems that common motion is not the only cue that is available to denote a single object. We see that the rod pieces are aligned, have common color and shape. Kellman and Spelke (1983) also investigated whether 4-month-olds perceive the unity of a stationary rod behind a block. Fig. 3.4, E2, shows the stimuli used in our model. They are analogous to those in the original study (not shown, compare Fig. 1.1). Infants were separated into three groups, one for each habituation display, respectively. After habituation each group was shown the test stimuli — a stationary "complete rod" and a stationary "broken rod".

In this study infants in the "rod occlusion" group did not discriminate between the test stimuli and dishabituated to them equally strongly. Why did infants show this behavior? Infants of the same age dishabituated clearly in the previous study with a moving rod. They were able to separate figure from ground, otherwise infants in the "complete rod control" group would have shown no preference. They were also able to distinguish the test stimuli because infants in the "broken rod control" group preferred the "complete rod".

S3: Original study

Another study with 4-month-olds was performed to clarify this issue, compare Fig. 3.4, E3. This time infants in the "rod occlusion" group dishabituated more to the "broken rod" test whereas the "rod pieces" group dishabituated more to the "complete rod" test.

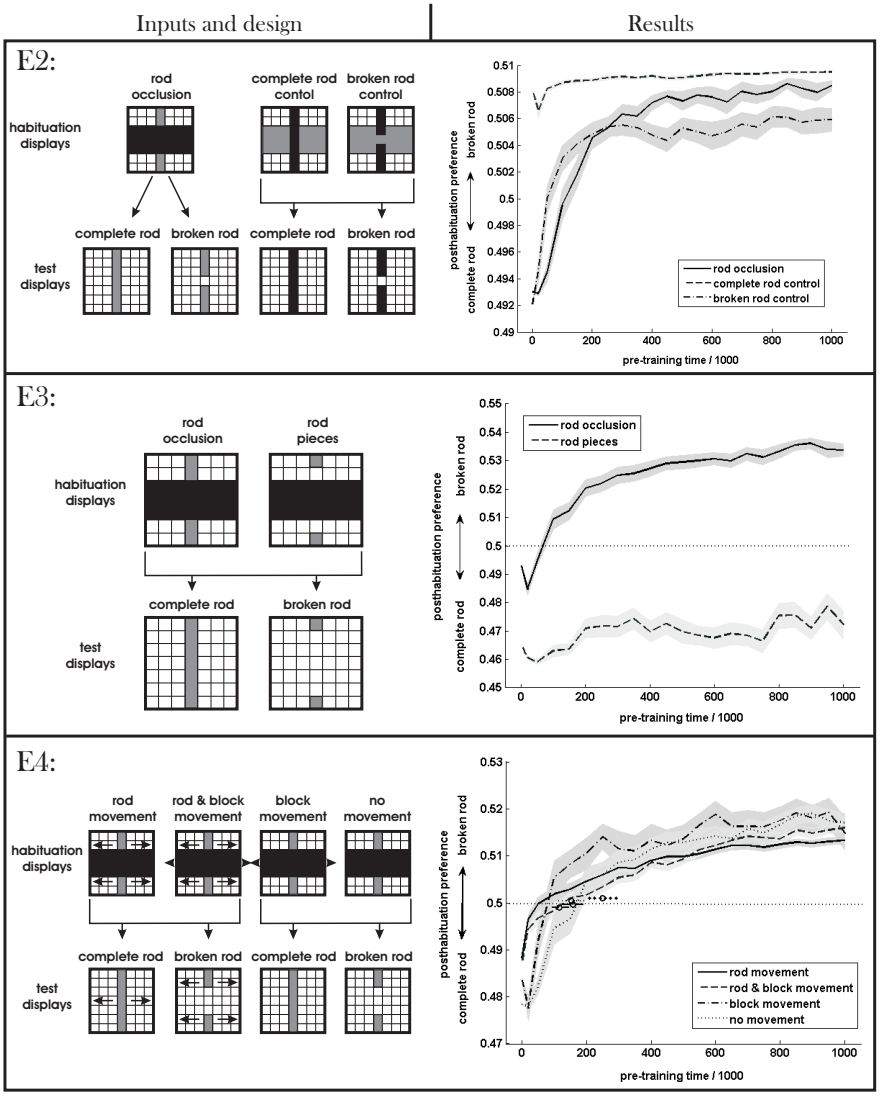

Figure 3.4: Left: input displays and design in Experiments 2, 3 and 4. Background (gray), foreground (black). Right: posthabituation preferences of the model in Experiments 2, 3 and 4. The gray shaded areas denote the standard error bars calculated on the basis of 25 simulations. In the results of E5, circles indicate where the corresponding preference curves intersect 0.5 and the lines through the circles are the respective error bars indicating to which precision this value is known. Note that these intersections are not where they seemingly should be according to the figure. See Sect. 3.8.4 for explanation.

In the light of this evidence, how can the "no preference" result for a stationary rod behind a block in Study 2 be interpreted? Suppose, infants perceive nothing at all behind the block or they perceive a different object other than a complete or a broken rod. If one of those hypotheses was true then the Study 3 should yield a "no preference" result as well. As just mentioned, this is not the case — infants prefer the "broken rod" in Study 3. Therefore, infants probably do perceive the rod ends in the "rod occlusion" display of Study 2 but have no definite perception of what is hidden.

Given the results of the last section, we know that 2-month-olds in the "rod movement" group did not show any preference. Therefore, we conjectured that infants at this age might be in a transition period for object unity perception. We hypothesize that in the case of the stationary rod occlusion, 4-month-olds may undergo a transition period as well since they do not show any preference. In the next study we will find an additional indication that this might be the case.

S4: Original study

In analogy to Fig. 3.4, E4, four groups of 4-month-old infants were assigned to each of the conditions and were habituated to them, respectively (Kellman and Spelke, 1983). After habituation the complete and broken rod stimuli were presented. Infants habituated to a moving rod were tested with moving rod displays and the other infant groups were tested with stationary rod displays.

Interestingly, only infants in the "rod movement" group showed a preference. They preferred the "broken rod" display which is consistent with the studies in the previous section. All other infant groups did not show any preference. Adults have been investigated as well by asking them to judge the strength of their impression of connectedness or disconnectedness of two visible parts of a display (Kellman and Spelke, 1983). In all four displays adults perceived the rod pieces as connected behind the block. Since adults perceive connectedness the "no preference" results can only be restricted to a limited age range. In other words, the infants must be in a transition period (assuming that the preference is a monotonic function of the age).

Given that on the one hand, infants are in a transition period for the stationary "no movement" display and on the other hand, they apparently have passed this period for the "rod movement" display this study supports the conjecture that the perception of object unity must be developing later for stationary rods as compared to moving ones.

Modeling procedure

The procedure is the same as in the previous section. The habituation and test stimuli are shown in Fig. 3.4.

E2: Results

In the "rod occlusion" case the development of the model's preference in Fig. 3.4, E2, is similar to the one in Experiment 1. Comparing this result to the original Study 2 we conclude that for a stationary rod & block display 4-month-olds must be in a transition period.

As for the "complete rod control" case both data of 4-month-olds as well as adult data (Kellman and Spelke, 1983) indicate a preference for the "broken rod" display. This is consistent with the model's results as well. The "broken rod control" result contradicts the experimental data since the Study 2 and adult data have shown that both 4-month-old and adults seem to prefer the "complete rod" display. See Sect. 3.5.3 for an explanation.

E3: Results

Similarly to the previous experiment the representation of a complete rod develops with time, Fig. 3.4, E3, "rod occlusion". As for the control condition ("rod pieces"), the model keeps preferring the "complete rod" throughout development. Both results are fully consistent with the results of Study 3 with 4-month-olds. Unfortunately, there is no data for newborns and adults to compare. Therefore, the model predicts that adults and newborns, people of all ages indeed, should

show a preference for the "complete rod".

E4: Results

Recall that only the "rod movement" group showed a preference for the "broken rod" display in the original Study 4. All other groups did not show any preference. On the other hand, adults interpret all of the habituation displays of Experiment 4 (see Fig. 3.4) as connected (Kellman and Spelke, 1983). Since adult data do not correspond to infant data we must conclude that the behavior of 4-month-old infants will change at some point later in development.

In Fig. 3.4, E4, we see the behavior of the model. For all conditions the preference starts below 0.5, transitions 0.5 at some point and saturates at a higher level in the end. Although the "rod movement" group does not undergo the transition before all other groups as has been observed with infants, the data are consistent with the observation that infants show no preference at some point (the area where the preferences cross 0.5) and that adults perceive the rod pieces as connected in each display.

This experiment confirms what we conjectured about moving vs. stationary rod perception since the network in the "no movement" condition crosses 0.5 significantly later than in all other conditions (see transition points in the results of E4, Fig. 3.4). Since there is only an indication from infant data that object unity might develop later for stationary objects as compared to moving ones (Study 4), the result of our model presents a prediction for future studies. See Experiment 5 for elaboration of this prediction and Sect. 3.5 for explanation.

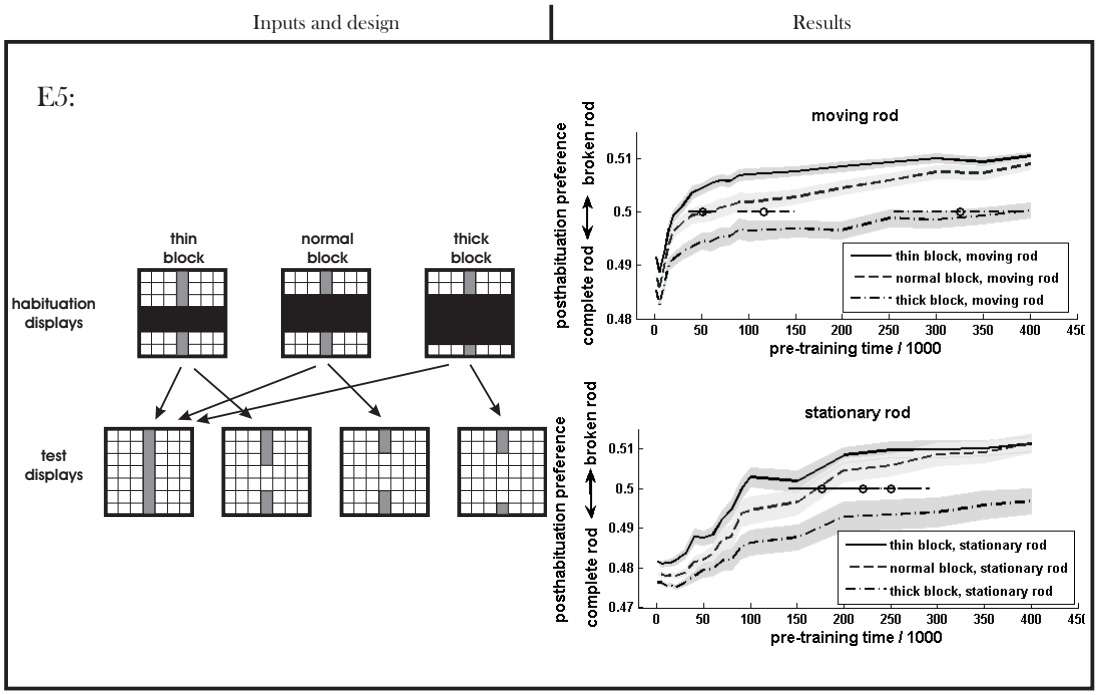

Figure 3.5: Left: input displays and design in Experiment 5. Background (gray), foreground (black). Right: posthabituation preferences of the model in Experiment 5. The gray shaded areas denote the standard error bars calculated on the basis of 25 simulations. Circles indicate where the corresponding preference curves intersect 0.5 and the lines through the circles are the respective error bars indicating to which precision this value is known. Note that these intersections are not where they seemingly should be according to the figure. See Sect. 3.8.4 for explanation.

3.3.3 Object unity perception behind occluders of varying sizes

S5: Original study

Johnson and Aslin (1995) and Johnson and Náñez (1995) made an object unity study just as in Study 1 but with 2-month-olds and stimuli presented on a flat computer screen (as opposed to the 3-dimensional stimuli in the original studies). Johnson and Náñez (1995) used the same width of the occluder as Kellman and Spelke (1983) but the 2-month-olds did not show any preference. This fits well to

the transition hypothesis formulated above. In another study (Johnson and Aslin, 1995) an occluder of half the width was presented and 2-month-olds *did* show a preference for the broken rod. This indicates that this transition point depends on the occluder width.

Modeling procedure

The modeling procedure is equal to Experiment 1 only that we now use the displays in Fig. 3.5 once in a moving rod condition and once in a stationary rod condition.

E5: Results

Fig. 3.5 shows the development of the model's preferences. First, the thinner the occluding block is the faster the model develops a perception of unity. Second, the preferences in a stationary rod condition develop later than the preferences in the corresponding moving rod condition.

The first result is in agreement with Johnson and Aslin (1995) whereas the second result makes the prediction that the perception of object unity should develop later for stationary objects as opposed to moving ones.

3.4 Perception of occluded object trajectories and object permanence

In the previous section we saw how the model explains how infants might learn to perceive the unity of objects. In this section we show that the model also learns to represent moving but temporarily invisible objects and thereby track their trajectories.

S6: Original studies

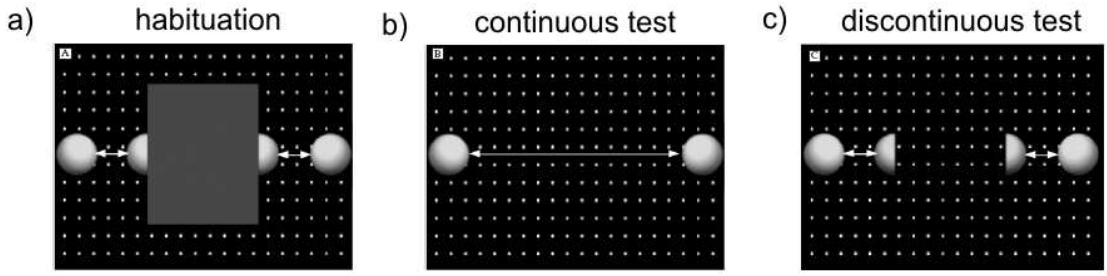

Figure 3.6: Design of the study. Infants were habituated with stimulus a) and tested with alternating stimuli b) and c).

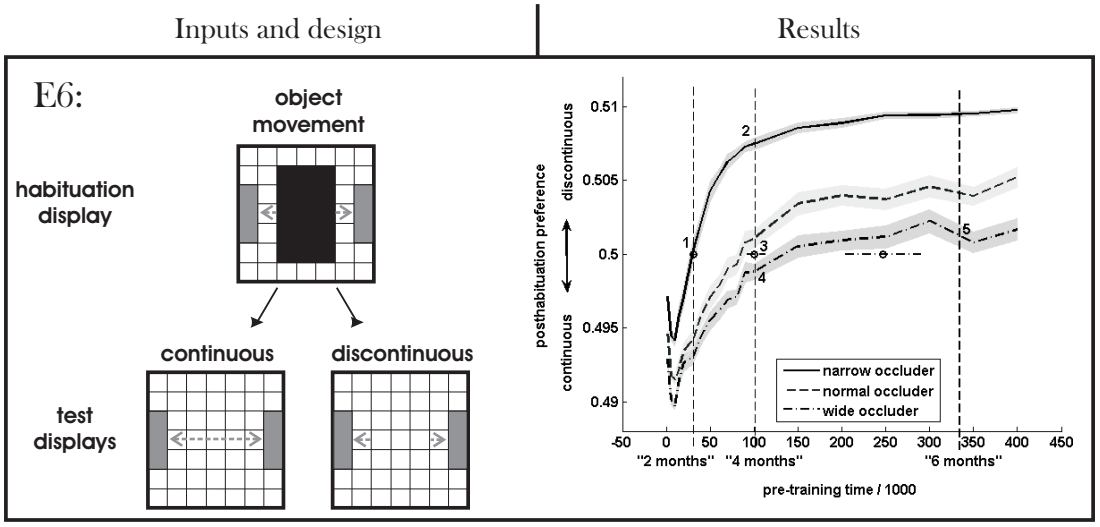

Figure 3.7: Left: input displays and design in Experiment 6. Background (gray), foreground (black). Right: posthabituation preferences of the model in Experiment 6. The gray shaded areas denote the standard error bars calculated on the basis of 25 simulations. Circles indicate where the corresponding preference curves intersect 0.5 and the lines through the circles are the respective error bars indicating to which precision this value is known. Note that these intersections are not where they seemingly should be according to the figure. See Sect. 3.8.4 for explanation.

Johnson et al. (2003) investigated whether and to what extent infants are able to track object trajectories behind occluders. Infants were presented a screen with a

ball oscillating behind an occluder. The ball would start on the left, move behind the rectangular occluder, reappear after some time at its right edge, move on, then reverse its moving direction, disappear behind the occluder and reappear again, see Fig. 3.6a).

Infant's interpretation of this habituation stimulus was measured by presenting two test displays. The continuous display, Fig. 3.6b), showed the ball moving fully visibly without being occluded at any point. The discontinuous display, Fig. 3.6c), showed the ball moving back and forth exactly as during habituation but disappearing at the locations where the occluder used to be.

The rationale behind this study was the following. If infants are able to track the ball trajectory behind the occluder during habituation they would generalize this to the continuous display and therefore dishabituate to the discontinuous display. If, on the other hand, the infants are not able to track the ball, they would only habituate to the directly visible parts of the trajectory which are identical to the discontinuous display. Thus, the infants would be more surprised to see the continuous display and dishabituate more to it. In a separate control study another group of infants was shown the test displays without prior habituation in order to assess any possible intrinsic preference.

The results were that infants' looking patterns depend on their age and on the width of the occluder. In case of a wide occluder 4-month-olds preferred the continuous display whereas 6-month-olds preferred the discontinuous one. Since the control study ruled out any intrinsic preference, this result is consistent with the interpretation that 4-month-olds are not able to track the ball behind the (wide) occluder while 6-month-olds have developed this ability.

A second study has been performed with a narrow occluder where the 4-month-olds suddenly preferred the discontinuous display. Thus, they were able to track the ball behind a narrow occluder. In consequence, just making the occluder more narrow allows infants to succeed in tracking. Maybe, in the case of the wide occluder, 4-month-olds partially succeeded in tracking the ball. In the same study 2-month-olds were habituated to a narrow occluder and did not show any pref-

erence. This result might mean that the tracking ability of 2-month-olds is just about to develop.

In a third study 4-month-olds were presented habituation displays with different occluder widths. Consistently, the wider the occluder became, the more the preference shifted towards the continuous display, i.e. the more difficult it seemed for the infants to track the hidden ball.

In summary, object tracking and permanence — the ability to track and represent things that are hidden behind other objects — seems to develop gradually in the first year of life. Moreover, this ability seems to develop more slowly for wide occluders as compared to narrow ones. This is plausible since objects disappear for a longer time behind a wider occluder.[3]

Modeling procedure

We modeled these studies in the same way as the previous ones. Note that the pre-training is never changed to adapt to these new input patterns. As habituation stimulus an object moving behind occluders of three different sizes is used (see Fig. 3.7, only the occluder with intermediate width is shown).

E6: Results

Fig. 3.7 shows the development of the network's performance for three different occluder widths. First note that for all widths the preferences grow above 0.5 at some point, i.e. the point at which the network learns to represent the moving hidden object. This point marks a qualitative change in the ability to represent occluded objects and is reached at different times in development depending on the width of the occluder. For the narrow occluder this point is reached first and it is reached last for the wide occluder.

[3]There is evidence that the width of an occluder and the time of disappearance expected by infants are correlated since infants seem to be able to represent the velocity of the moving object (von Hofsten et al., 2007).

Taken together this explains the pattern of data of all three studies. In Fig. 3.7 the first study corresponds to the intersections 4 and 5 (4- and 6-month-olds were presented a wide occluder). The second study corresponds to the intersections 1 and 2 (2- and 4-month-olds were presented a narrow occluder). And finally the third study corresponds to the intersections 2, 3 and 4 (4-month-olds were presented occluders with three different widths). Thus, the observed pattern matches the behavioral pattern of infants.

3.5 A unified account of the development of object unity, object permanence, and occlusion perception

The model provides a concrete picture of how the perception of object unity, permanence, and tracking might develop.

3.5.1 Learning object unity

At the start of the pre-training the network has not developed any representations yet. When presented with the habituation stimulus, it learns only what is directly visible. Fig. 3.8 shows the mechanism for the case of the "rod movement" display in the object unity experiments. In the beginning of learning the background hidden layer learns to represent the separate rod pieces (broken rod) and the foreground hidden layer learns to represent the block. Since the network represents the broken rod, it "dishabituates" only to the "complete rod" test display which explains the behavior of newborn infants (Slater et al., 1990).

As pre-training proceeds, the network learns to represent a complete rod since non-occluded, complete rods are abundant during pre-training. Specifically, the network learns that two rod pieces moving together are always *correlated* with the presence of a connection between those pieces, i.e. a complete rod. This mechanism is similar to pattern completion in auto-encoder networks (Hertz et al.,

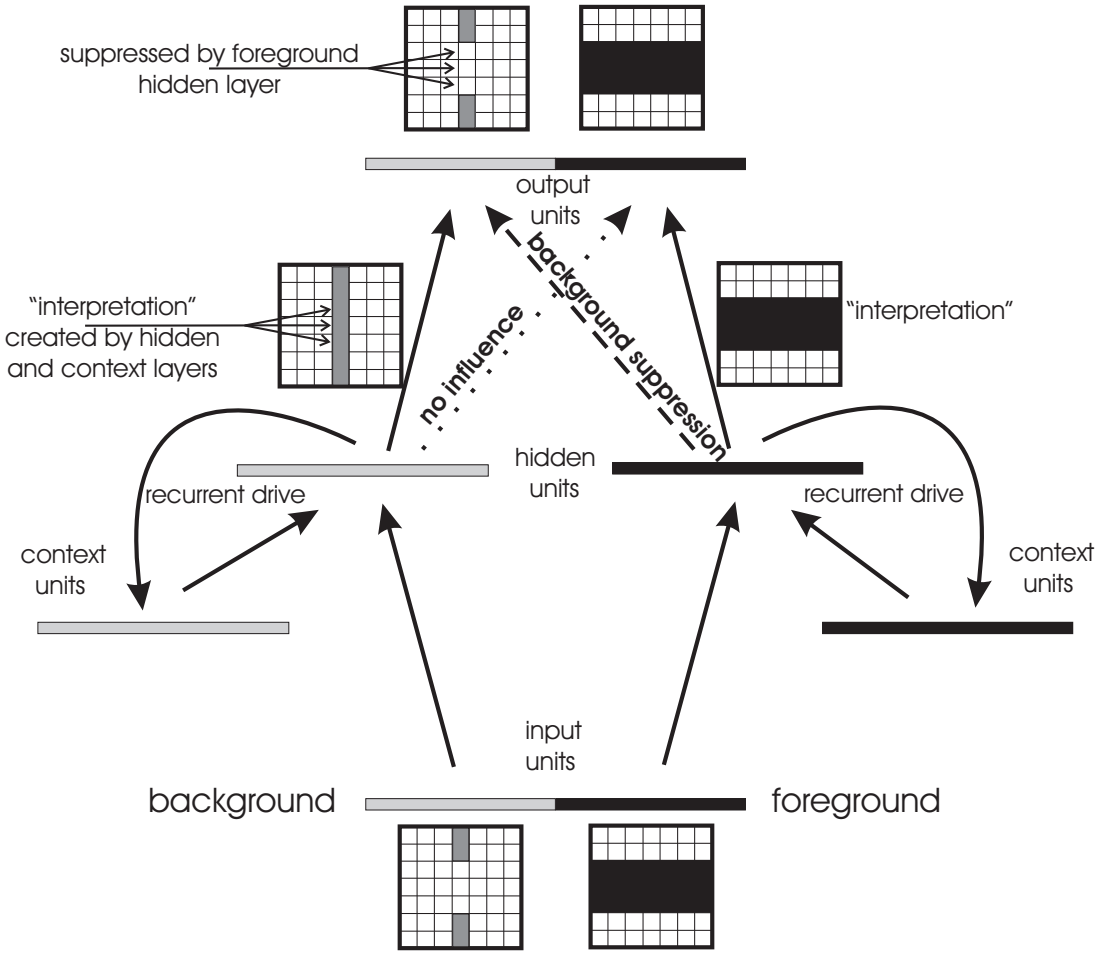

Figure 3.8: Object unity and permanence representation in the network

1991). Therefore, when presented two rod pieces moving together, the network automatically retrieves the representation of the intermediate piece, i.e. it "fills in" a complete rod, see "interpretation created by the hidden and context layers" in Fig. 3.8. At the same time the network learns to present a broken rod at the output layer through "background suppression" (strong negative weights) such that the output matches the input at the next time step (a broken rod) while at the same time the background hidden layer represents a complete rod. In the performance measurement, this representation is compared to test displays. Therefore, the comparison to the broken rod leads to a greater difference than the comparison to the complete rod, i.e. the network "dishabituates" to the "broken

rod". Since learning is a continuous process in the network, there will also be a transition period where the network shows no preference for either test display as supported by experimental evidence (Johnson and Náñez, 1995). In that case, the representation of a complete rod is only "half-built".

In the case of a thin occluder there is already a lot of rod surface directly visible. Therefore, there is a strong "correlation drive" in the background hidden layer to "fill in" the missing part. On the other hand, if only little rod surface is present then the correlation drive is weak and the area to be "filled in" is large. Thus, it takes more learning time in order to develop a strong correlation drive for thick occluders. That explains the slower learning for thicker occluders in Experiment 5 as supported by Study 5.

If the visible surface is disrupted as in the "rod pieces" habituation display of Experiment 3, the "filling in" of this gap is suppressed by the learning algorithm during habituation. This is due to the fact that the correlation drive can "fill in" only the occluded parts of the display, otherwise it would cause a prediction error which leads to a suppression of the "filled in" gap. That is why the network always represents just the rod pieces and therefore shows a preference for the complete rod at all pre-training times explaining the results of Study 3.

In summary, the model learns to perceive object unity based on a statistical correlation of the presence of object parts.

3.5.2 Learning object permanence and tracking

Similar to above, the network begins without any representation before the pre-training. Therefore, when presented the habituation display in Fig. 3.7, it learns only what is directly visible, which corresponds to the discontinuous test display. This explains the model's preference for the continuous test display after small pre-training times, explaining the results of Study 6 for young infants.

As pre-training proceeds, the network learns to represent a free, non-occluded motion since they are abundant during pre-training. When some background object

becomes occluded the background input layer becomes deprived of input. Since the network is recurrent, it is the *recurrent drive* that ensures that the representation is kept in the hidden layer for several time steps, which explains object permanence. It also explains object tracking since the representation that is maintained is not necessarily static but represents moving objects dynamically.

The recurrent drive can not represent an occluded object for an arbitrary amount of time without being supported by direct input. The representation becomes weaker with time until it is lost completely. This explains the occluder thickness dependence of Study 6, since objects have to be tracked for several time steps if the occluder is thick.

In summary, the model learns to track objects and to represent them behind occluders because of the developing recurrent drive.

3.5.3 Role of edge configurations

Recall that the "broken rod control" result in Experiment 2 is not consistent with the experimental data (Kellman and Spelke, 1983). 4-month-olds and adults both seem to perceive a broken rod in front of a block, see Study 2. As for the model, it represents the central part of the rod whenever two aligned rod pieces are presented. This is due to the fact that the model is not sensitive to edges, since in the "rod occlusion" display the edges of the rod and the occluder form a T-junction which is an important cue revealing the depth relation between the rod and the occluder while the edge configuration in the "broken rod control" case indicates that the rod pieces are in front of the occluder.

According to Kellman and Arterberry (1998), pp. 160–161, the sensitivity to edge configurations (which they call edge-sensitive process) has not developed before the first 6 months of the infant's life. With regard to Study 2 this is probably not the case. If the configuration of edges in the habituation displays was meaningless infants may treat the "rod occlusion" and the "broken rod control" displays equally which is not supported by evidence. Therefore, this study presents a hint that the edge configuration has some influence.

In our view, the picture that is most consistent with the data is the following. Recall that 4-month-olds show no definite perception of the occluded part in the "rod occlusion" display (Study 2). It is not possible that they just perceive the directly visible rod ends because that would lead them to dishabituate more at the "complete rod" display comparable to neonates' behavior. Therefore, direct, uninterpreted perception always leads to the perception of a broken rod in these displays, i.e. a definite empty space between the rod ends. As the infants did not show any preference, they have probably taken into account the edges and interpreted the space between the rod ends as at least indefinite, which is consistent with the experimenters' conclusion. Thus, the consideration of the edge configuration had the *opposite* influence on the looking behavior as compared to direct perception which explains the "no preference" result. Considering the "broken rod control" display both direct perception and the edge configuration may tell the infant that the rod is broken, since the edge relations lead them to infer that the rod pieces are indeed in front of the block. Therefore, the infants showed a preference for the "complete rod".

Our model can not take into account edge configurations and therefore it is not surprising that it treats the "rod occlusion" and the "broken rod control" similarly. On the other hand, the model also demonstrates that the sensitivity to edge configurations is not necessary for the interpretation of a complete rod even for stationary displays. It seems only necessary for the inference of the correct depth relations, at least for the object unity tasks.

3.5.4 Object perception of neonates?

Slater et al. (1990) conclude from their studies that (1) infants are able to segregate figure from ground and (2) treat the two rod ends as separate objects. We incorporated the first conclusion as an assumption of our model. Regarding the second conclusion, an alternative interpretation is simply that newborn infants have direct, uninterpreted perception. This is how our network starts learning and it is consistent with the result of the study since infants could just have habituated to the directly perceived rod ends and then dishabituated to the complete rod.

3.5.5 Object unity for stationary vs. moving objects

An analysis of the network reveals that during the habituation phase the representation of a "filled-in" complete rod behind an occluder tends to get erased. This is due to the absence of the connection of the rod ends in the input and the continued learning of the network during habituation which may erase the representation of the connection in the background hidden layer, see Fig. 3.8. This effect is much stronger for a stationary rod than for a moving one due to the fact that for a moving rod the connection moves as well, i.e. it appears at many positions. Since it is harder to erase a representation that has been built up at many positions as compared to the representation at only one position the representation of a moving rod turns out to be more stable. Consequently, it takes more time for the network to develop a stable representation of a stationary occluded rod which explains its behavior in Experiment 5.

This result constitutes a prediction of our model that the perception of object unity develops later for stationary objects as compared to moving ones. This assumes though, that infants do not track the occluded object smoothly, otherwise the moving rod would appear stationary in eye-centered coordinates and not differ from a stationary rod. This possibility is unlikely though as suggested by experimental data. For example, in an object tracking study von Hofsten et al. (2007) observed that "none of the [ten 4-month-old] infants persued the object smoothly over the occluder", p. 636, and "the gaze stopped at the occluder edge until making saccadic shift(s) over to the other side", p. 632.

3.6 Predictions of the model

A good model should not only be able to explain experimental data but also make specific and testable predictions, i.e. it must be falsifiable. In the following section all predictions of our model are summarized.

1. The main prediction of the model is that infants can **learn** the perception of object unity, permanence, and tracking. Specifically, the model predicts that

visual preferences of infants at different ages are a result of their visual experience. By changing the amount and structure of their visual input one can alter their visual processing. Specifically, the natural development of preferences might be accelerated, slowed down or even reversed. For example, 2-month-olds who did not show any preference after the "rod movement" habituation, see Study 1, could be taught to develop a preference for the "complete rod" display — the way neonates behave — by presenting them disconnected objects in common motion. It is also conceivable that infants growing up in a world governed by different statistics than our world might not learn object unity at all, e.g. they might learn that parts that move together actually belong to different objects and vice versa.

2. The perception of object unity develops later for stationary occluded objects as compared to moving ones, see Experiment 5.

3. Slater et al. (1990) concluded that neonates interpret the "rod movement" display in Fig. 1.1 as two separate rod pieces. In contrast, our model predicts that newborn infants do not perceive objects at all (see Sect. 3.5.4) which is also in agreement with the studies of Slater et al. (1990). Specifically, neonates only have an uninterpreted, bottom-up perception of the display.

4. The perception of object unity is corrupted if there is a visible gap between the object part and the occluder, see "rod pieces" display in Experiment 3, Fig. 3.4. Kellman and Spelke (1983) have shown that already for 4-month-olds but according to our results (Experiment 3), the model predicts that it should hold true for all ages.

5. Graded dishabituation time: The dishabituation time depends on the amount of surface in the test display that matched the expected surface. For example, if in Experiment 1 the "broken rod" display had shown more rod surface (if the gap between the rod ends had been smaller), then infants who already perceive unity will dishabituate less to it as compared to the display with a

larger gap.

3.7 Discussion

We presented a model that can explain a broad range of infant studies (12 studies in total from three different laboratories). The guiding principle for learning is the prediction of future inputs (Elman, 1990). The network is able to learn to represent the unity of objects as well as object permanence and tracking behind occluders thereby providing a unifying account of these different event categories. The model specifically explains the gradual development of these abilities, encompassing the behavior from newborn infants to adults. It also accounts for specific effects like the role of the width of occluders and differences in the perception of moving vs. stationary objects. Finally, the model demonstrated that these abilities can be acquired through learning and made predictions about future experiments.

3.7.1 Related modeling work

Schlesinger and Young (2003) used a prediction network for modeling tracking of objects behind occluders (Baillargeon, 1986). It was presented stimuli similar to the original infant studies but they did not pre-train their network which we consider essential in this kind of modeling (see Sect. 3.2.2 for details). It is important since infant behavior undergoes qualitative changes during development and therefore every developmental model needs to be able to model the developmental processes that give rise to different competence levels at different ages. Therefore, we have to provide the model with some age dependent pre-experimental experience which we accomplish by pre-training our network.

To our knowledge, the first attempt to analogize the error in neural networks to infant looking time has been made by Mareschal et al. (2000). We adopted a modified version of the network error that is suitable for our network structure comparing the representations by the background and foreground hidden layers with the inputs as defined in Sect. 3.8.2.

Mareschal et al. (1999) trained a network to predict occlusion events but the network comprises specialized modules trained differently for their specific tasks (e.g. an object recognition module) which makes it less parsimonious than our model. Also, the trajectory prediction module learns to predict the position of a specific object which is only a subset of the whole retinal input (that includes the occluding screen as well). That makes the model less general and difficult to extend to additional studies.

Similarly to our model Munakata et al. (1997) trained a simple recurrent network to build representations of occluded objects but their model was trained exactly for the one task is was supposed to accomplish: representing an occluded object at a fixed position whereas in our model neither the position of the objects nor their size nor the sort of task is pre-specified.

An important predecessor of this model is the work done by Mareschal and Johnson (2002) on object unity perception who presented a neural network model successfully explaining how the perception of object unity may be learned. Our model extends and compliments this line of work in several respects. First, we provide a more detailed modeling of the actual experimental studies on object unity. Second, Munakata and Stedron (2002) pointed out that a potential shortcoming is the usage of a supervised target signal training the network's response to classify each input as "single object", "two disjoint objects" or "intermediate" which begs the question of whether such a target signal is available to infants. Our work addresses this question by extracting the target output directly from the environment. This constitutes a generic principle of prediction of future inputs (Elman, 1990) that is supported by neuroscientific evidence (Rao and Ballard, 1999). Third, our generic target output allows for modeling the habituation paradigm more closely which allows the construction of modeling variables (dishabituation error) that correspond to actually measured looking time in habituation studies. Fourth, although the construction of a number of pre-programmed task-specific modules was carefully justified by the existence of corresponding innate abilities in infancy, our model uses fewer such assumptions and predefined programming. Together with our generic target signal it makes our model more extendable and generalizable which

is attested by its ability to account for additional visual events beyond object unity such as the perception of object permanence and occluded object trajectories.

Finally, the main difference to all previous work is that we presented a unified model that can account for infant behavior in several domains, not only in the domain of object unity. We consider this as major strength in modeling work specifically because such models are likely to reveal general principles that guide development. We are confident that we will be able to extend the model to represent even more visual events, like launching (already successful, see Chapter 2), blocked motion, gravity and support, continuity of object motion, object identity (see discussion of future work below). Up to now, technical reasons hinder us from doing so (long training times for large networks trained with backpropagation of error, convergence to local minima of the error landscape).

3.7.2 Nativist perspectives

Our model provides a mostly empiricist account of the development of object unity and permanence. The only major assumption that we made is that a simple form of figure-ground segregation is present at birth, which is supported by experimental evidence (Slater et al., 1990). Many of the modeled studies are regarded as evidence for innate capacities though (Kellman and Arterberry, 1998, Spelke, 1990, 1994) which is questionable in our view. As argued by Spencer et al. (2009) "nativists routinely extrapolate well beyond the data, making bold claims about time points not directly under investigation. For instance, Marcus (2001) described a habituation study with 4-month-olds, concluding 'it seems likely that at least some of the machinery that infants use in this task is innate' (p. 370), but he presents no evidence to support this claim." In our case, Spelke extrapolates similarly beyond the data for 4-month-olds in order to support claims about innateness although four months are more than enough time for learning given that infants are able to learn effectively even within two minutes (Saffran et al., 1996).

As for object unity perception, Kellman and Arterberry (1998), pp. 158–160, argue that there are unlearned foundations. Since there is a clear development in the perception of object unity taking place between birth and four months of age

they postulate the existence of the edge-insensitive process, which is supposed to be responsible for the detection of common motion, and that maturates at the age of two months. Moreover, they argue that given the evidence that newborns seem to perceive the rod ends as separate objects (Slater et al., 1990), the perception of object unity could not be learned because it would force infants to start with a wrong perceptual rule, e.g. "things that translate rigidly are separate". If the learning account was true, a learning process would need to overwrite this rule which they consider as implausible.

Although we agree that this particular learning account is implausible there is no evidence that there should be a perceptual *rule* for newborns. Neonates could just have uninterpreted perception. In fact, our model *demonstrates* that it is possible to start with uninterpreted perception, apply a learning algorithm to it and learn to perceive object unity without the assumptions that Kellman and Arterberry (1998) considered necessary. Apart from that our learning account is more parsimonious since it does not assume a special-purpose edge-insensitive process with an arbitrary time of onset.

More general arguments have been put forward by Spelke (1990, 1994). According to this view initial core principles such as perception of object unity could not be learned because "learning systems require perceptual systems that parse the world appropriately" (Spelke, 1994), p. 439, i.e. if the perceptual systems can not parse the world appropriately into objects then no learning about object properties could be possible. Spelke assumes that in order to learn one needs to be able to reason about some entities. Therefore, those entities need to be present before learning can start. Gestalt psychologists refer to this as the paradox of the "experience error" (Spelke, 1990). In contrast to these considerations, our model *demonstrates* that learning is possible without those entities (like objects). In the end, the model learns almost from scratch and arrives at the same capacities that infants develop.

3.7.3 Limitations of the model and future work

The main limitation of the model is of a technical nature: since learning in such a large network is difficult with backpropagation of error, the range of possible training sequences had to be limited, e.g. the rod could only have width equal to 1 and be vertically centered in the display, see Sect. 3.2.2 for a full description. If the network did not have these constraints, we could have just trained it with rectangles with random heights and widths that move around and occlude each other once in a while. In an earlier version of the model this approach was already successful but it had biologically implausible and somewhat arbitrary constructions that made that model less parsimonious.

Another limitation is that we are not able to explain all occlusion data. Especially, data reviewed by Baillargeon (1999) poses a great challenge. Specifically, it is not clear why the behavior of infants seems to depend on whether the parts of the occluding screen are connected or what is the role of the size of occluded objects. Some of the reviewed data indicates that infants might first acquire a general *behind/not behind* distinction and later in development add more details to these concepts. This kind of data cannot be explained within the current framework and will be discussed in the next chapter.

Another interesting topic would be using this framework to account for different event categories. The framework is suitable for these challenges since it is not constructed specifically for the object unity or occlusion tasks. In Chapter 2 we have already demonstrated that a model of this type is able to represent both occlusion and launching events. Therefore, except for technical reasons the extension to launching is straightforward in our model. If launching is possible then visual events such as blocked motion, motion continuity, object identity, gravity and support and even containment are also conceivable. There is a huge amount of experimental data available on these topics and much of it may be explainable by the present framework. It definitely poses a great and exciting challenge for future work.

3.8 Model details and equations

3.8.1 Calculating the neuron activities

The network consisted of $N_z = 7 \times 7 \times 6 = 294$ input and output units, respectively. Each hidden and context layer had 50 units. According to Fig. 3.1 the network calculated its activity in the following way. The hidden layer activities, $\vec{Y}^1(t)$ and $\vec{Y}^2(t)$, resulted from

$$\vec{Y}^k(t) = s\left(V^k\,\vec{X}^k(t) + U^k\,\vec{Y}^k(t-1)\right), \tag{3.2}$$

where $k \in \{1, 2\}$, V^k and U^k are weight matrices and s is the sigmoid

$$s(x) = (1 + e^{-x})^{-1}. \tag{3.3}$$

The output, $\vec{Z}(t)$, is calculated accordingly:

$$\vec{Z}(t) = s\left(W^1\,\vec{Y}^1(t) + W^2\,\vec{Y}^2(t)\right), \tag{3.4}$$

with the W^k being weight matrices to the output layer.

3.8.2 Learning: backpropagation through time

The network in Fig. 3.1 is defined by the equations in the previous section. We used the backpropagation though time algorithm (Hertz et al., 1991) that reduces the squared difference between the output vector, $\vec{Z}(t)$, and the next input vector, $\vec{X}(t+1)$:

$$E(t) = \frac{1}{2}\sum_{i=1}^{N_z}(Z_i(t) - X_i(t+1))^2. \tag{3.5}$$

We want to know how the connection weights U^k_{mn} have to change in order to reduce the objective function (3.5), for those are the weights that are involved in the recurrent calculations. Since the calculation for U^1_{mn} and U^2_{mn} is the same, we will treat them as the same in the calculation and omit the index k. The gradient

of the objective function is

$$\frac{\partial E(t)}{\partial U_{mn}} = \sum_i (Z_i(t) - X_i(t-1)) \, \frac{\partial Z_i(t)}{\partial U_{mn}}. \tag{3.6}$$

From (3.4) and (3.2) it follows

$$\frac{\partial Z_i(t)}{\partial U_{mn}} = Z_i(t) \, (1 - Z_i(t)) \, \sum_j W_{ji} \frac{\partial Y_j(t)}{\partial U_{mn}} = Z_i(t) \, (1 - Z_i(t)) \times$$

$$W_{ni} \, Y_n(t) \, (1 - Y_n(t)) \, \underbrace{\left[Y_m(t-1) + \sum_{j,k} U_{kj} \frac{\partial Y_k(t-1)}{\partial U_{mn}} \right]}_{Q_{mn}}. \tag{3.7}$$

Expanding the last term leads to

$$Q_{mn} = Y_m(t-1) + \left(\sum_j U_{nj} \right) \, Y_n(t-1) \, (1 - Y_n(t-1)) \times$$

$$\left[Y_m(t-2) + \sum_{l,k} U_{lk} \frac{\partial Y_l(t-2)}{\partial U_{mn}} \right]. \tag{3.8}$$

Of course this expanding never ends since we are in a loop and can in principle go infinitely back in time. Practically, we must cancel the second term in the square bracket. Canceling this term in eq. (3.7) leads to the simple Elman network (Elman, 1990) that does not involve backpropagation through time. Canceling this term in eq. (3.8) leads to going back one step in time. This change improved the network's performance considerably.

After calculating the gradient, at each time step we changed the weights, U_{nm}, using eq. (3.6):

$$\Delta U_{nm} = -\frac{\partial E(t)}{\partial U_{nm}}. \tag{3.9}$$

3.8.3 Performance measurement in the network

In analogy to what might happen in infants we compare the networks representation of a e.g. complete rod during habituation to the test displays. In order

to do so we have to retrieve this information from the hidden layers. In Fig. 3.8 this representation is described as "interpretation created by hidden and context layers". This "interpretation" can be calculated by

$$\vec{Z}^1_{\text{int}}(t) = s\left(W^1\,\vec{Y}^1(t)\right) \tag{3.10}$$

in the case of the background hidden layer. Note that in comparison to the real output calculation, eq. (3.4), the influence of the foreground hidden layer $W^2\vec{Y}^2(t)$ is omitted. Given the networks "interpreted" output, $\vec{Z}^1_{\text{int}}(t)$, how can its performance be measured and related to an infant's looking time? For this purpose, the difference between the "interpreted" output during habituation with the actual inputs, $\vec{X}^{(1)}(t)$ and $\vec{X}^{(2)}(t)$, during the tests, has been computed. This results in what we call the dishabituation error, in analogy to the dishabituation time:

$$E_k \equiv \sum_{a=1}^{2} \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{N_z} \left| Z^a_{i,\,\text{int}}(t) - X^{(k)}_i(t) \right|, \tag{3.11}$$

where $\vec{Z}^2_{\text{int}}(t)$ is given by

$$\vec{Z}^2_{\text{int}}(t) = s\left(W^2\,\vec{Y}^2(t)\right), \tag{3.12}$$

analogously to (3.10) and $T = 16$ is the duration of each trial. After computing these entities we obtain the model's preference using (3.1). Note that the whole network is symmetric and the performance measurement is also symmetric in the foreground and background calculations.

3.8.4 Calculating points of intersection and the respective error bars

In the results of Fig. 3.7, for example, the intersection points of the preference curves with 0.5 had to be calculated. In order to judge whether two preference curves cross 0.5 at significantly different times, an estimation of the mean point of intersection and the corresponding standard errors are necessary. For this purpose, a 6th grade polynomial is fitted to a preference curve and the intersection point

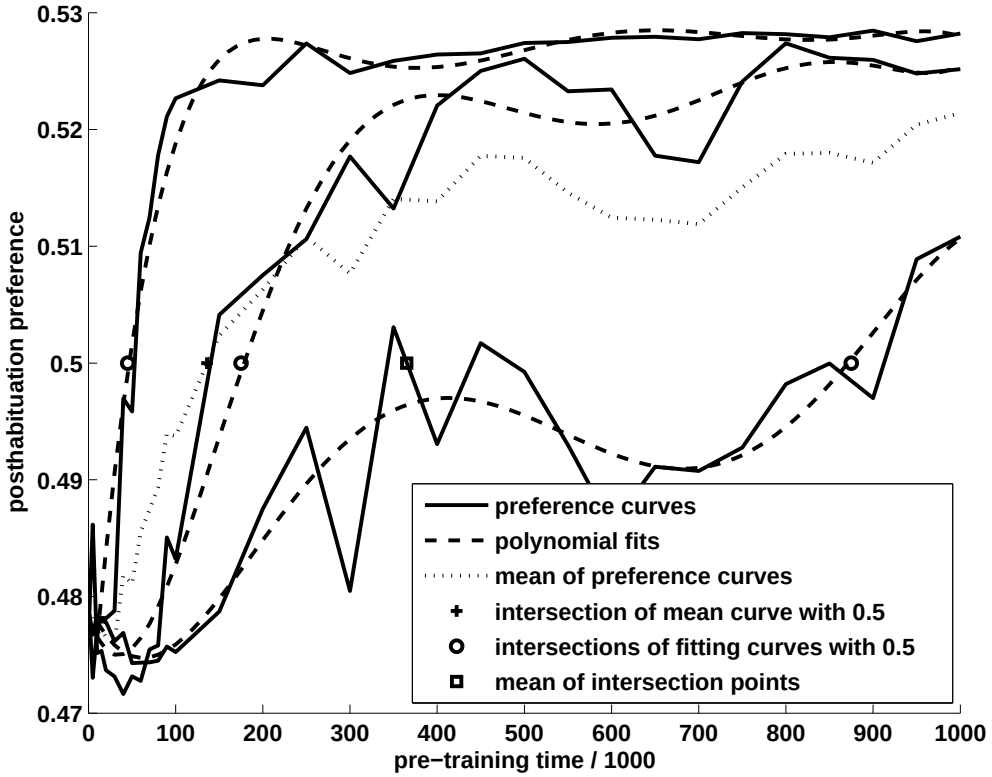

Figure 3.9: Illustration of the calculation of intersection points. The preference curves shown are part of the real data.

of the polynomial with 0.5 is calculated. 5th grade polynomials have been tried as well which did not make any difference in the results. Subsequently, the mean and standard error of the set of those intersection points could be calculated. This resulted in the means and error bars in Fig. 3.7.

Fig. 3.9 shows how the intersection points are calculated. Specifically, note that the intersection of the mean curve with 0.5 is different from the mean of the intersection points. At first glance this might seem surprising but for each set of non-linear curves this is not avoidable because the distribution of intersection points depends strongly on the shape of the curves.

Chapter 4

Evaluation of progress

How should the progress made by the models be evaluated? Initially, the work was motivated by finding and modeling the first steps in the development of cognition. Specifically, we set the goal of modeling how infants segregate the world according to the role that its entities play in it. One starting point was the animate vs. inanimate distinction occurring early in life and proposedly based on the perception of self-propelled vs. caused motion.

Modeling the study by Leslie (1982) seemed like a prospective project to embark on this problem since it included caused motion (the "direct launching" condition) and self-propelled motion (the "no prior movement" condition). Since the model was trained to represent and predict launching scenarios it was not surprising that self-propelled motion caused a higher prediction error than caused motion. This fact alone may guide us to the conclusion that the model has learned the animate vs. inanimate distinction in a primitive sense. Such conclusions should be taken with care though as will be discussed below.

Intuitively, we as adults have the impression that there is some categorical difference between caused motion and self-propelled motion. It is important to ask: what is the nature of the prediction error made by the model in the "no prior movement" test? The model predicts the second block to stay immobile since it is not pushed by the first block. But, contrary to this prediction, it does move and causes a prediction error at two pixels: (1) at the pixel where it was expected to

stay, but where it did not stay and (2) at the pixel where it was not expected to arrive but where it did arrive. Thus, the error's magnitude is two pixels, i.e. the error depends on the resolution of the model. Generally, our criticism aims to show that the prediction error will depend on many characteristics, e.g. the resolution, the form, size and velocity of the involved objects etc. Therefore, it would be premature to conclude that the model has acquired the distinction between animate vs. inanimate. In other words, it is able to show smoothly varying prediction errors but it has not acquired a categorical distinction between self-propelled and caused motion. In the following, we will elaborate on this first intuition, showing that the model on object unity and occlusion suffers from this problem as well. Specifically, we will discuss a set of experimental data supporting our intuition and which can not be captured by the discussed models even in principle. This will lead the discussion of the question that maybe, something important is missing.

4.1 Challenging data

4.1.1 Object unity

The general rationale behind the object unity studies discussed above is the following. If infants represent a partly occluded rod as connected they should be surprised seeing them disconnected in the test phase. Although these results and its development during infancy have been modeled successfully in the previous chapter, it would be premature to conclude that the model is able to represent unity or connectedness as such. Intuitively, the notion of connectedness or unity does not vary with the distance that separates the parts of an object. Placing pieces of an object, like the two rod ends, more distant from each other, does not make them more disconnected. The rod still is disconnected even if the gap between the two rod ends is very small. Intuitively, infants should react to the connectedness rather than to the size of the gap because only the strict connectedness of two rod ends makes their correlated motion likely. We suspect that infants' understanding of this issue makes them behave in the observed way. Specifically, we conjecture that prediction 5 (suggesting that infants' dishabitation time depends

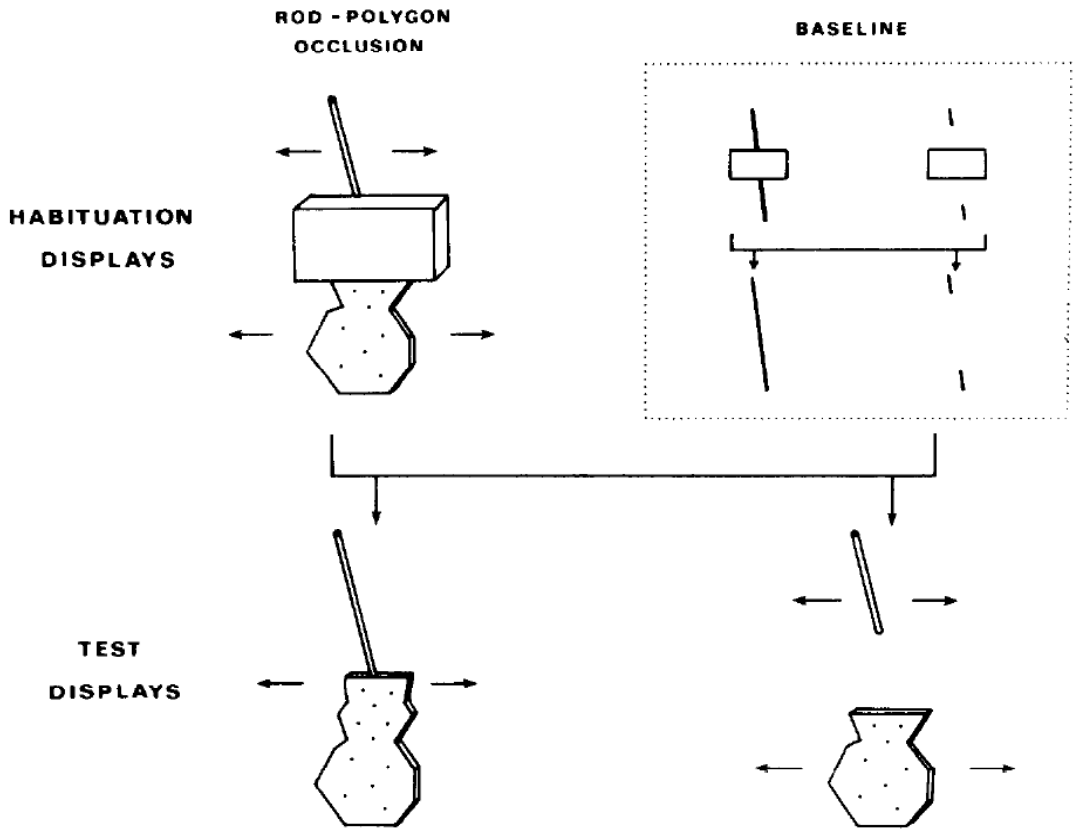

Figure 4.1: Displays and design in Experiment 6 by Kellman and Spelke (1983).

on the size of this very gap) may be wrong. Experiment 6 in Kellman and Spelke (1983) supports this conjecture. In this study the lower rod end was replaced by an irregularly shaped polygon that moved coherently with the upper rod end (Fig. 4.1). Although never exposed to such stimuli infants readily dishabituated when presented the rod end as disconnected with the polygon. This study has turned to be difficult to account for in our model since the representation of unity or connectedness relies heavily on the model's previous experience (pre-training) with unoccluded objects that would reoccur occluded in the experiments. Since a polygon never occurred during pre-training the model generally "dishabituated" with this stimulus. In a nutshell, previous experience seems to be important in order to generally grasp that parts moving together are usually connected while abstracting away from specific forms and features. Similarly, one could argue that the model for causality is essentially on the right track, it just did not succeed in

abstracting away from the specific forms of the blocks, the resolution, velocities etc.

Despite the attractiveness of this view, we have to recall the experimental data discussed in the introduction. There, infant studies indicated that global categories are formed first and only later the categories are subdivided into more specific and detailed subsets. The process is not a process of abstracting away from specific to general, from fine to coarse, but one from general to specific, from coarse to fine. In this context, it is fruitful to discuss some more data concerning these issues.

4.1.2 Baillargeon's data on causality, occlusion, support and containment

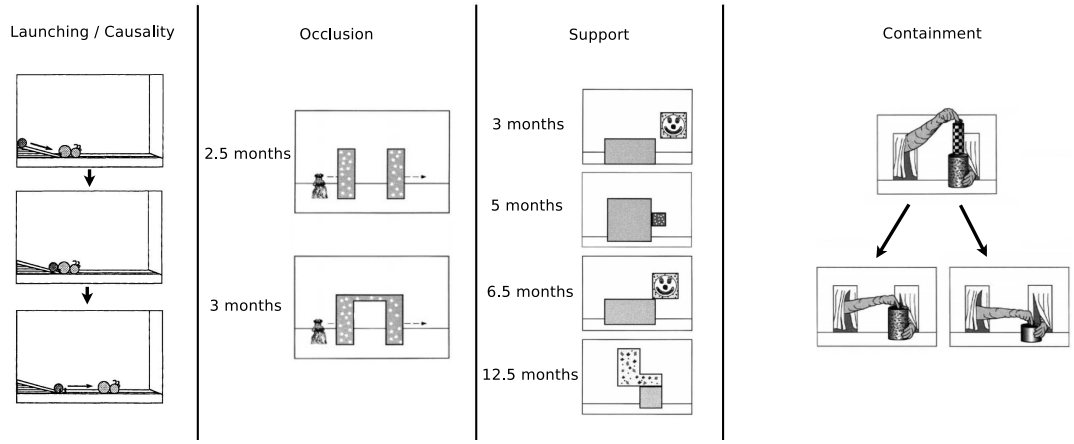

Figure 4.2: Stimuli in Baillargeon's studies on causality, occlusion, support and containment.

In her review papers (Baillargeon, 1994, 1999) Renée Baillargeon discusses a wide range of experimental data that support the coarse-to-fine hypothesis for the perception of events (rather than just the development of categorization that we discussed in the introduction). It is not possible to go into detail of these studies here but we summarize the main results.

In the case of launching and causality, infants at the age of already 2.5 months

can judge that an object will only move when launched by another object and that it will not do so when not caused to (Fig. 4.2, launching/causality). On the contrary, not until around 6 months they start to understand the role of the sizes of the objects and how they influence the velocity and trajectory of the objects. It seems that infants first acquire a *launched/not launched* distinction and only later are able to judge the details of the scenario.

In the case of occlusion it was observed that at the age of 2.5 months infants dishabituate if a mouse moving behind two parallel bars does not occur between them but they do not dishabituate if those bars are connected by a thin bar (Fig. 4.2, occlusion). In other words, infants at 2.5 months readily detect occlusion violations with two parallel bars but they fail to do so when presented a U-shaped object even though the mouse should be visible in the interior area of the "U". Only around 3.5 months they detect the occlusion violation for the U-shaped object. Baillargeon suggests that infants acquire first a *behind/not behind* distinction leading them to expect the mouse being hidden when behind the U-shaped object. Only later they learn to adjust their expectations to the shape of the object as well.

Of course, there is more to event perception than we could cover in this thesis. Specifically, support and containment are among the most studied events in object perception. As for support events, Baillargeon (1999) studied how the perception of objects supporting each other develops. As shown in Fig. 4.2, support, infants at the age of 3 months seem to consider mere contact of two objects sufficient for support as they are not surprised when the object is in contact with the side of another object while they are surprised to see an object supported without contact. As infants grow older (5 months) mere contact does not seem to be sufficient for support any more but contact at the upper surface does the job even though the contact surface may not be large enough. Only by 6.5 months infants learn to be sensitive to the amount of contact surface and by 12.5 months they succeed in judging the weight distribution of the supported surface. This data suggests that the perception of support starts also with a binary *contact/no contact* distinction and gets refined later, analogously to previously discussed event categories.

As for the perception of containment experimental data revealed that only at

the age of around 6-7 months infants start to consider the width and the height of an object that is lowered into a container, e.g. that a tall object cannot be lowered into a small container (Hespos and Baillargeon, 2001), see Fig. 4.2, containment. On the other hand, 3-month-olds are already sensitive to whether the container is opened or closed, e.g. that objects cannot be put into a closed container. Again, it seems that infants first acquire a binary *opened/closed* distinction before they consider the role of detailed parameter relations.

Looking back at object unity data by Kellman and Spelke (1983) in the light of this data it seems that infants have grasped that connectedness is essential for object unity, i.e. they acquired a *connected/disconnected* distinction early in their development.

Taken together, the data suggests that infants initially start with understanding something like the "essence" of an event. They start to understand *that* something is launched to move but not yet how far it will move. They understand *that* an object is behind another but not yet where or how much of it is occluded. They understand *that* there is contact but not yet where and how much of contact is necessary for support. Finally, they understand *that* two parts are connected but not how precisely they are connected. Looking at this data, it seems that infants somehow succeed at grasping the "essence" of an event first while leaving out all the unimportant details. How could they arrive at this knowledge?

Given the discussed data it looks like infants are able to distinguish features of an event and learn them step by step. For example, in the case of causality they first learn a launched/not launched distinction and only later consider the sizes of the objects involved. In all cases it seems that infants first grasp the essential feature of the event (whatever that means) and later add more and more features to the phenomenon.

4.1.3 Comparison to model performance

Comparing that to our models leads to the observation that the models behave differently. For example, in no training phase does the model in the last chapter acquire a connected/disconnected distinction with disregard to the object shape even though the model has been trained with rectangles of various sizes. Nowhere in the hidden units there is a representation of a connected/united object. Similar criticism applies to the representations of occlusion and causality.

Despite the successes of the models in explaining a wide range of data, the previous discussion leads to the conclusion that something important may be missing. How is it possible to build a model that develops representations of the events in a featurewise fashion? Specifically, how is it possible to extract such features of the visual input in an unsupervised way? Note that nobody tells infants to first extract the "unity"-feature of objects. The model achieves that by learning correlations of inputs when they occur unoccluded but these correlations are defined by the shape of the object that the model is trained with. Thus, there is no shape or size independence of the unity representation.

This issue in itself does not pose a substantial problem but it indicates that it is part of bigger problem: the object representation is not divided in several features that are acquired step by step in development. For example, the movement of rods of different lengths are learned together in the model although it may be more efficient to possess a representation of say "left motion" combined with the representation of "rod of length x". As research indicates (Baillargeon, 1999) infants acquire the various variables like "rod length" separately from other features.

This leads us to the important question: How do infants know what features to extract? Obviously, once the features are given it is possible to build effective representations of stimuli, e.g. representing the length of a rod separately from its position. The crucial question is not how to extract those features once they are known but how to find them.

4.2 Conclusion and further steps

There is a lot of material on feature extraction in the machine learning literature. One path for further research may be finding a mathematical structure for feature extraction and continue to model the featurewise development of event perception in infancy. Although this procedure promises to be fruitful in explaining more experimental data, it does not answer the question how *infants* extract those features, why they prefer extracting them over some other features and why some features are extracted later than others. Intuitively, it seems that infants extract the "essential" features first as has been argued above but we still do not know why some features are preferred over others and what "essential" means precisely. In order to answer these questions it seems necessary to investigate how infants generally begin to grasp regularities in the environment. How do infants come up with ideas about anything at all? Be it an effective representation of unity or launching or anything else. One of the research guidelines of this thesis includes staying very close to experimental data. Therefore, it was decided to study very basic mechanisms of how infants begin to understand regularities and structure in their visual environment. The last chapter will deal with these questions and model how infants generally may form expectations based on their understanding of a visual scene.

Chapter 5

Development of visual expectations and sequence learning

5.1 Introduction

The formation of visual expectations is important for infants' understanding of spatiotemporal events, planning and predicting sequences of actions, an indicator for information processing and even a predictor for adult IQ (Haith et al., 1988, Canfield et al., 1997, Benson et al., 1993, DiLalla et al., 1990). It is usually studied by the Visual Expectation Paradigm (VExP, Haith et al., 1988) where infants are presented left and right appearing image sequences while eye movement reaction time is measured and classified as anticipatory or reactive. Numerous studies have shown that even young infants are able to form visual expectations revealed by a drop in reaction time and higher probability of anticipatory eye movements toward upcoming stimuli (Canfield et al., 1997, Wentworth and Haith, 1998, Jacobson et al., 1992). Moreover, studies with more complex stimulus sequences have revealed that infants can cope even with sequences requiring the ability to enumerate repeatedly displayed stimuli at the same position (Canfield and Haith, 1991, Canfield and Smith, 1996, Rose et al., 2002, Reznick et al., 2000).

Despite these results it is unclear why infants' reaction time drops in the course

of a session (Haith et al., 1988, Wentworth and Haith, 1998) and of development (Canfield et al., 1997, Rose et al., 2002). Furthermore, why infants perform anticipatory eye movements at all remains a topic of controversy and speculation. For example, Haith et al. (1988) speculate that this behavior "serves to maintain continuity in an ever-changing perceptual world" or "give[s] rise to more cognitively based planning skills". Canfield and Haith (1991) suggest that an infant tries to "gain internal control over its behavior, with the result that actions can be executed more smoothly and stably in a complex and changing environment". Already Piaget noted that "... this anticipatory function ... is to be found over and over again at every level of the cognitive mechanisms and at the very heart of the most elementary habits, even of perception" (Piaget, 1971, p. 19). Despite the importance of this phenomenon there exist only vague speculations about the origins of the observed behavior and a universally accepted theoretical account is still lacking. Specifically, this phenomenon connects to the question posed in the last chapter: how do infants start to grasp simple regularities in the environment such as a sequence of alternating images?

In this section we address this issue and propose a theory for the development of visual expectations in infancy. We show that the theory accounts for a wide set of experimental data. In essence, the theory suggests that infants strive to maximize their looking time at each interesting stimulus, which is only possible if the structure of the sequence is understood.

In order to formulate the theory more precisely we implement it as a computational model - a recurrent neural network model developed by Lazar et al. (2009) that forms internal representations of input sequences. We proceed by extending it by a reinforcement learning architecture that learns to perform anticipatory eye movements. In contrast to classical neural network modelling approaches our model uses local learning rules and contains several biologically plausible elements like excitatory and inhibitory spiking neurons, spike-timing dependent plasticity (STDP), intrinsic plasticity (IP) and synaptic scaling.

The model allows us to account for twelve experimental studies on visual expectations from four different laboratories. Furthermore, it is shown how infants'

internal representations can change even if the input remains stationary, which allows infants to enumerate repeated occurrences of the same stimulus and to predict its shift to a different position successfully. In this way our model sheds fresh light on (sequential) subitizing processes in infancy. It also makes new predictions for future studies, e.g. that relearning of a sequence should happen faster than learning it for the first time.

5.2 Review of experimental data

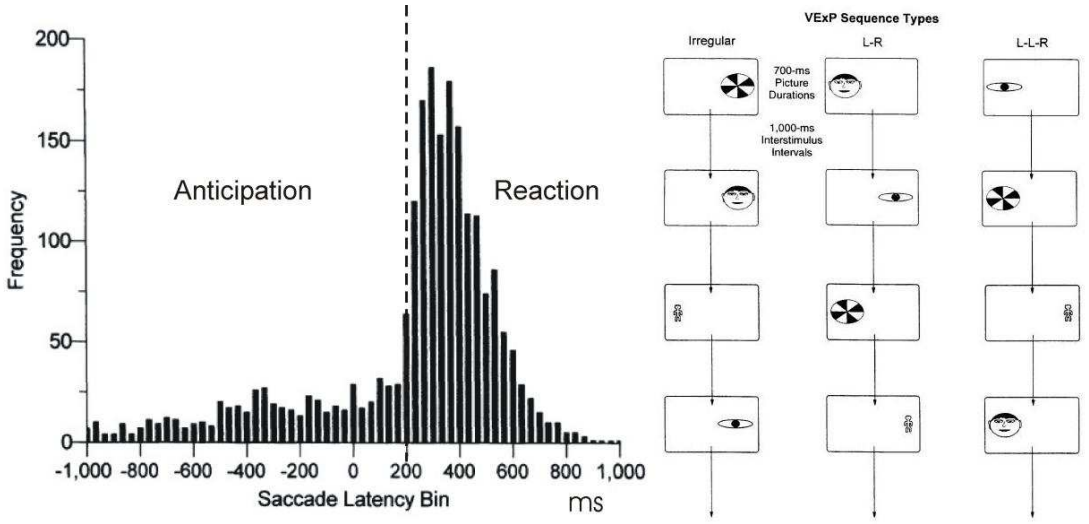

Figure 5.1: Left: saccade latencies in a Visual Expectations Paradigm. Right: Experimental design of VExP studies. Adopted from Canfield et al. (1997).

The Visual Expectation Paradigm (VExP) developed by Haith and his colleagues (Haith et al., 1988) permits to study how infants use spatiotemporal regularity to forecast upcoming events. Infants watch brief sequences of computer-generated pictures that are either regular (e.g. pictures appear in a deterministic fashion on the left and right sides of the monitor) or irregular (see Fig. 5.1, right). After each onset of a stimulus the reaction time is measured, defined as the time difference between the beginning of a gaze shift towards the stimulus and its appearance on the stimulus. A gaze shift is usually classified as anticipatory if it began after the

offset of the previous stimulus and before 200 ms after the onset of the subsequent stimulus taking into account that infants are estimated to require at least 200 ms for a reaction (compare Fig. 5.1, left). In this way the percentage of anticipatory saccades is calculated.

In this section we review general experimental results that ask for explanations. They can be summarized in the following way.

1. For regular sequences, with increasing age and trial number,
   a) the reaction time decreases,
   b) the probability of correct anticipations increases while the probability of wrong anticipations decreases.

For the left-right (LR), left-left-right (LLR) and left-left-left-right (LLLR) sequences,

2. as sequences get more complex[1]
   a) the reaction time increases
   b) the differences between reaction times toward sequences of different complexity increases with age,
   c) the probability of correct anticipatory gaze shifts to the right decreases,
3. the probability of an anticipatory gaze shift to the right increases with each presentation of a left stimulus.

For the so-called pivot sequence (left-top-left-bottom, see below),

4. the probability of correct anticipations is higher for the pivot stimulus (left) than for either of the wing stimuli (top and bottom).

[1] "complexity" refers to the length of the sequence in this case.

Result 1) has been observed both as an age effect and as a within-study effect. The fact that the reaction time decreases as infants grow older (1a) is strongly supported by a number of both longitudinal and cross-sectional studies (Canfield et al., 1997, Reznick et al., 2000, Rose et al., 2002, Canfield et al., 1995, Jacobson et al., 1992). One part of the explanation is certainly that infants' reaction time generally decreases with age even for unpredictable sequences. In spite of that a number of studies succeeded in proving that infants actually develop spatiotemporal expectations in the course of the study, see e.g. Wentworth and Haith (1998), Haith et al. (1988), observing a decrease in reaction time with the number of trials. These results are consistent with the interpretation that infants learn to represent spatiotemporal sequences.

As for result 1b), Canfield and Haith (1991) observed that 3-month-olds are more likely to perform an anticipatory gaze shift from the first/second/third L to R (in the LR, LLR and LLLR sequences, respectively) and from R back to L than 2-month-olds. Rose et al. (2002) observed that the percentage of anticipations in the RRL sequence increased significantly from 15.7% (5-month-olds) to 16.9% (7-month-olds) to 24.5% (12-month-olds). Consistent results hold for left-right alternating studies (Reznick et al., 2000, Jacobson et al., 1992).

Result 2a) was obtained by Canfield and Haith (1991), Fig. 3 who observed a linear increase in reaction time of 3-month-olds as the sequences changed from LR, LLR, LLLR to IR (irregular). Additionally, this effect is smaller for younger infants (2-month-olds) (2b). In the same study the authors observed that the probability of anticipating the right stimulus decreased for more complex sequences (2c).

Result 3) was obtained by Canfield and Haith (1991) and Canfield and Smith (1996) who found that the probability to shift the gaze to the right increases with each presentation of a left stimulus in the LLR and LLLR sequences. It was concluded that infants form number-based expectations related to enumeration processes.

Reznick et al. (2000) studied a sequence where the stimuli can occur at three possible positions: left (L), top (T) and bottom (B). The stimuli were displayed

in a deterministic manner: L-T-L-B-L-T-L-B.... Apart from the results already stated the authors observed that the probability of an anticipatory gaze shift toward that pivot stimulus was almost twice as high as toward either wing stimulus (result 4) for all ages tested (6-, 9- and 12-month-olds).

What all these studies have in common is the lack of a unified explanation for their results although specific speculations about the underlying reasons for the results are discussed. In our view, this explanation together with the observed results constitute the requirements for a good theory of visual expectations in infancy.

5.3 A theory of visual expectations

Our theory is based on the following five principles.

P1 Infants' try to maximize the looking time at a stimulus.[2]

This principle explains why infants react to upcoming stimuli at all after seeing an empty screen. However, since it takes around 200 ms to execute a saccade towards a stimulus, the time spent looking directly at the stimulus can be further increased by starting to move the eyes to its location even before it appears. Therefore, infants will try to make predictions based on regularities in the sequence. Hence, the decrease of reaction time and increase of anticipatory eye movements will go hand in hand and can be explained by an increased ability to predict the sequence. This addresses the question about why infants make anticipatory eye movements at all.

This leads us directly to the next question: how do infants learn the structure of a sequence and make predictions?

P2 Each input causes a different pattern of neural activations in the infant brain.

[2]For justifications of this view the reader is referred to the general discussion is Sect. 5.9.

P3 The learning of sequence representations is mediated by a plasticity mechanism causally linking those patterns.

Imagine a population of neurons in the brain of an infant. At each point in time some neurons will be active (i.e. emit a spike) and others will be non-active. This activity pattern will change from one point in time to the next. Now, suppose the infant is exposed to the pivot sequence: L-T-L-B-L-T-L-B... The principle P2 suggests that each input will cause a different pattern S_i in the brain as shown in Fig. 5.2A since visual exposure to a stimulus activates photoreceptors in the retina which in turn influence the activity in the whole brain.

Now that a sequence of activity patterns is created, the plasticity mechanism postulated by P3 links the patterns causally, i.e. makes pattern S_i *cause* S_{i+1} even in the absence of input (see, Fig. 5.2B). Therefore, it is possible to say that the brain has learned the sequence and is able to predict future inputs.

Although P2 and P3 establish a framework how the brain can learn a sequence it is not clear how it could capture the structure of a sequence. Mapping each input to a different pattern does not enable the brain to capture statistical regularities in the sequence - an ability that infants certainly possess. Further postulates are necessary.

At this point we make the following definition:
Let $I_1I_2I_3$... be a sequence of inputs. A **minimal required memory (MRM)** of an input I is the length of the shortest sequence preceding I and determining it uniquely.

For illustration consider the sequence LTLBLTLB... What is the MRM of B? Immediately before B there is an L which does not determine B uniquely since an L could also be followed by a T. Therefore, a longer history needs to be considered. Does the sequence part TL determine B? This time the answer is yes, since every time there is a T followed by an L we can be sure that a B will follow (TL $\rightarrow$ B). Thus, the MRM of B is 2. Similarly, the MRM of T is 2 as well (BL $\rightarrow$ T). As for L, it is always determined by either a single T or a single B. Therefore, the MRM

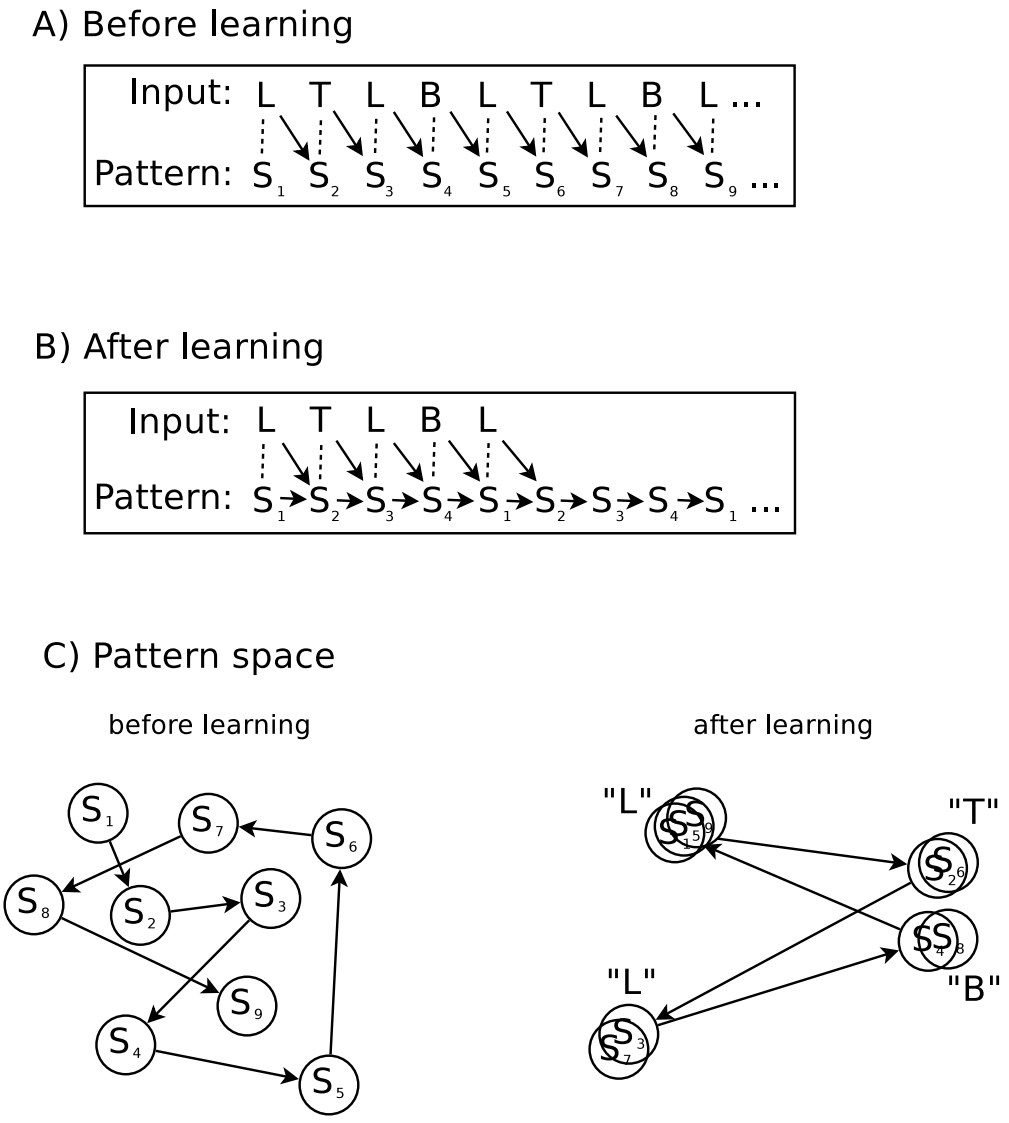

Figure 5.2: **A**: The displayed input sequence causes various patterns in the infant brain denoted by solid arrows. **B:** A plasticity mechanism makes the patterns cause each other such that the brain is able to "predict" the next pattern / stimulus. **C:** Before learning the patterns are random. After learning patterns are reorganized according to principle P4.

of L is 1.

This definition is now applied in the next principle:

P4 Infants' ability to predict a future input decreases as the input's MRM grows

larger.

Intuitively, the MRM is linked to the amount of working memory needed in order to predict the next input. In other words, it is important how many inputs of a sequence have to be kept in mind in order to be able to predict the next input. If this number is large, the prediction task is harder and the performance will suffer.

Applying P4 to the results summarized above turns out to be very fruitful. As just discussed, for the pivot sequence it is easier to predict the L than the T or B. Therefore, more correct anticipations can be expected for L as confirmed by result 4). Long sequences like LLLR certainly require more memory than shorter ones. Thus, predicting them is more difficult which accounts for result 2). Similarly, in the LLLR sequence the prediction of each L and finally R grows harder since its MRM grows larger: R $\rightarrow$ 1$^{\text{st}}$ L, RL $\rightarrow$ 2$^{\text{nd}}$ L, RLL $\rightarrow$ 3$^{\text{rd}}$ L, LLL $\rightarrow$ R. Therefore, the anticipation of R becomes more probable as the sequence proceeds which accounts for result 3). Taken together, the principles P2, P3 and P4 account for how sequences are learned.

Since P2 and P3 are expressed in terms of brain activity, it is fruitful to do the same for P4 as well. The formed activity patterns in the brain, as postulated by P2 have to be read out and transformed into an eye movement in order to produce observable behavior including predictive eye movements. If a large MRM makes the prediction task more difficult, two patterns whose MRM is large will be similar since distinguishing similar patterns is more difficult (see Fig. 5.2C, e.g. "T" and "B"). A conclusion from P2 and P4 is therefore that patterns with large MRM and similar history will be harder to distinguish and therefore more difficult to predict. This effect also explains why the patterns S_1, S_5, S_9 etc. are mapped very closely to each other: a long history has to be kept in mind in order to distinguish the kth from the $k+1$st occurrence of the sequence block LTLB. In fact, as we shall see below, the patterns become so similar that is it reasonable to assign the same name S_1 to them as we did in Fig. 5.2B.

In order to clarify what happens during development and during a study we suggest the fifth principle:

P5 During development as well as during a study, the activity patterns corresponding to the sequence inputs become more distinct.

P5 accounts for result 1) since as patterns grow more distinct the read-out can be performed more easily leading to a higher percentage of correct anticipations and smaller reaction time.

Taken together the five principles account for the various experimental results and also suggest the underlying causes for these results which are difficult to address experimentally. In the following we suggest a computational model that embodies these principles and proves that a mechanistic implementation is possible. Furthermore, the behavior of the model matches qualitatively experimental results described above.

5.4 Computational model

In order to implement a computational model of the outlined theory of visual expectations it is necessary to use a mathematical structure that is capable of representing and predicting an input time series. Recurrent neural networks (RNNs) are natural candidates for this task and will thus be quickly reviewed in the next secion (see Lukosevicius and Jaeger (2009) for a more elaborate review).

5.4.1 Recurrent neural networks (RNNs)

The characteristic feature that distinguishes RNNs from feedforward neural networks is that the connection topology processes cycles as opposed to simply computing a function. Therefore, RNNs are able to develop a self-sustained temporal activation dynamics even in the absence of input and to form a dynamical memory that is able to process temporal information.

Formally, the problem to be solved by RNNs can be defined as a problem of learning a functional relation between a given input $\vec{U}(t) \in \mathbb{R}^{N^U}$ and a desired output $\vec{Z}_{\text{target}}(t) \in \mathbb{R}^{N^Z}$, where $t = 1, ..., T$, and is the number of data points

in the training dataset $\{(\vec{U}(t), \vec{Z}_{\text{target}}(t))\}$. In supervised learning paradigms the relation is learned through minimization of some error measure $E(\vec{Z}, \vec{Z}_{\text{target}})$. On the other hand, in unsupervised learning paradigms the target output is not available, therefore the network parameters are updated to optimize some measure defined on the network state and input only.

RNNs have been shown to be universal approximators of dynamical systems under fairly mild assumptions (Funahashi and Nakamura, 1993) and therefore appear as highly promising tools for time series processing. Despite these advantages and their natural correspondence to neural networks in real nervous systems the impact of RNNs in nonlinear modeling has remained limited for a long time. This is mainly due to substantial difficulties in training such networks. For example, gradient-descent methods for minimizing the error measure can drive the network dynamics through bifurcations (Doya, 1992) and therefore convergence can not be guaranteed. Further, updates can be computationally expensive since many update cycles may be necessary leading to long training times.

In this difficult situation, in 2001 a fundamentally new approach to RNN design was proposed under the name of *Liquid State Machines* (Maass et al., 2002) and *Echo State Networks* (Jaeger, 2001) collectively referred to as reservoir computing. In this paradigm the RNN is *randomly* created and remains unchanged during training. The RNN is called *reservoir* and is passively excited by the input signal and maintains in its state a nonlinear transformation of the input history. The desired output signal is generated as a linear combination of the neuron signals from the input-excited reservoir. This linear combination is obtained through linear regression, using the teacher signal as a target. The general idea behind such designs is using the reservoir to expand the input history $\vec{U}(t), \vec{U}(t-1), ...$ into a rich enough reservoir state space $\vec{X}(t) \in \mathbb{R}^{N^E}$, while the output neurons combine the neuron signals $\vec{X}(t)$ into the desired output signal $\vec{Z}_{\text{target}}(t)$. Therefore, the "purpose" of the reservoir is to contain a rich enough, high-dimensional representation to make that possible.

The resulting systems could outperform previous methods in a wide range of tasks and reservoir computing are therefore established methods nowadays (Lukosevi-

cius and Jaeger, 2009). Despite these successes it is unlikely that a randomly designed and static reservoir is optimal for more complex tasks. Therefore, research in reservoir computing proceeds to explore various training mechanisms for the reservoir. In this section we will make use the self-organizing recurrent neural network developed by Lazar et al. (2009) that used spike timing dependent plasticity and intrinsic plasticity to shape the reservoir structure.

5.4.2 Model architecture

The computational model[3] is based on a self-organizing recurrent neural network Lazar et al. (2009) and extends it by a reinforcement learning architecture. Its construction is based on the principles suggested above.

Recurrent reservoir

The network reservoir consists of N^E excitatory (E) and $N^I = 0.2 \times N^E$ inhibitory (I) threshold units (Fig. 5.3)[4]. Neurons are connected through weighted synaptic connections, where W_{ij} is the connection strength from unit j to unit i. All possible $E \rightarrow I$ and $I \rightarrow E$ connections are present, while the $E \rightarrow E$ connections are random and sparse. Self-connections are prohibited.

At each time step, a subset of N^U input units and another subset of N^U gaze units receives positive drive $U = 1$ when their state is active. The input units and gaze units encode the position of the current visual stimuli and the current gaze position, respectively. For example, the network can be presented a left-right alternating input sequence by alternatingly activating the input units as shown in Fig. 5.3 which could roughly be related to different regions of the retina of an

[3] MATLAB code is available upon request.

[4] An active excitatory/inhibitory neuron enhances/inhibits the depolarization (activation) of all neurons it projects to.

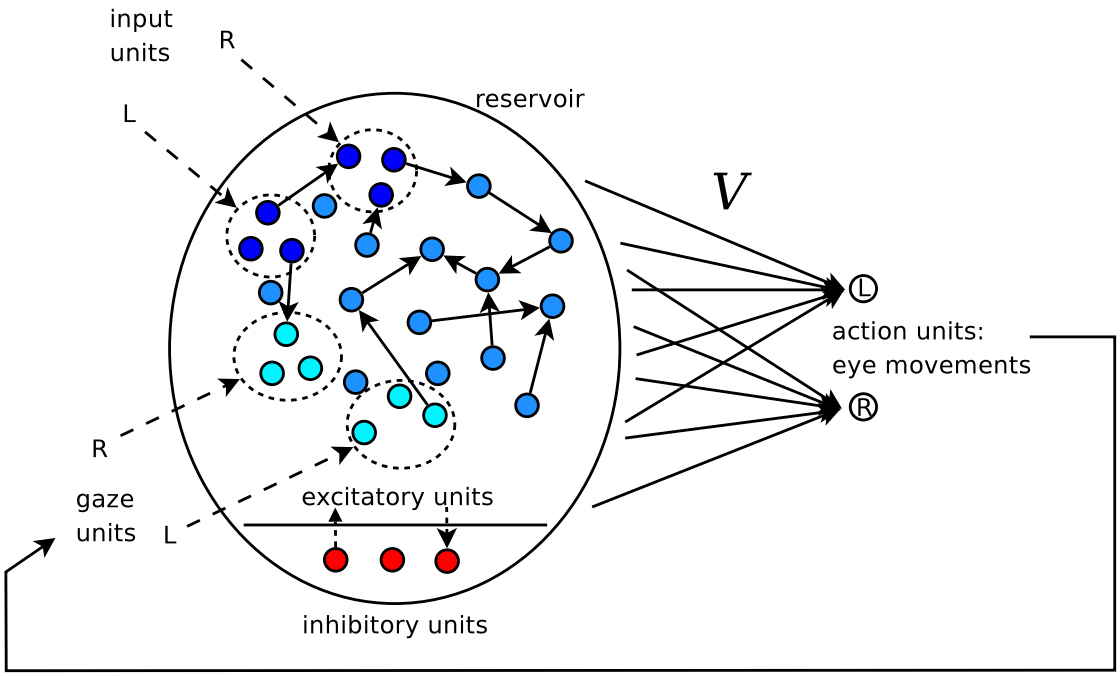

Figure 5.3: The network architecture. Input sequences drive the recurrent reservoir. The reinforcement learning architecture performs eye movements that change the gaze position in the next time step. This gaze position is provided as additional input to the network. Goal: make eye movements such that the gaze position matches the input.

infant watching an alternating stimulus sequence. The gaze units carry oculomotor feedback representing the current looking direction of the eyes of the infant. Activity in the input and gaze units spreads through the recurrent network and leads to a series of activity patterns. This firing activity of the network at the discrete time t is given by:

$$X_i(t+1) = \Theta\left(\sum_{j=1}^{N^E} W_{ij}^{EE}(t)\, X_j(t) - \sum_{j=1}^{N^I} W_{ij}^{EI}(t)\, Y_j(t) + U_i(t) - T_i^E(t)\right) \tag{5.1}$$

and

$$Y_i(t+1) = \Theta\left(\sum_{j=1}^{N^E} W_{ij}^{IE}(t)\, X_j(t) - T_i^I\right), \tag{5.2}$$

where $\vec{X} \in \mathbb{R}^{N^E}$ and $\vec{Y} \in \mathbb{R}^{N^I}$ are activations of excitatory and inhibitory units, respectively, W_{ij}^{EE}, W_{ij}^{EI}, W_{ij}^{IE} are synaptic weights, $\vec{T}^E$ and $\vec{T}^I$ are threshold values, Θ is the Heaviside step function, and $\vec{U}$ is the additional input drive received only by a predefined selection of input and gaze neurons (see dashed ellipses in Fig. 5.3). Basically, the network is driven by the input and gaze units and maps this activity into a high-dimensional reservoir space. Such a network fulfills principle P2: input sequences will drive the whole network population showing activity patterns.

Principle P3 demands that a plasticity mechanism imprint the causal structure of the sequence into the network. Therefore, following Lazar et al. (2009) the network is equipped with a biologically inspired learning rule: spike timing dependent plasticity (Bi and Poo, 1998, Markram et al., 1997). In their simplified version a connection strength between a pre- and a postsynaptic neuron is increased if the postsynaptic neuron fired a spike one time step after the presynaptic neuron and it is decreased if the timing is reversed:

$$\Delta W_{ij}(t) = \eta_{\mathrm{STDP}} \left(X_i(t)\, X_j(t-1) - X_i(t-1)\, X_j(t) \right). \tag{5.3}$$

In this way, two patterns that *happen* to succeed each other are causally linked together such that the first pattern *makes* the second pattern occur at the next time step.

The network is also equipped with other biologically inspired mechanisms such as intrinsic plasticity: the firing threshold of a neuron adjusts such that it keeps its average output firing rate constant at the value H_0:

$$T_i^E(t+1) = T_i^E(t) + \eta_{\mathrm{IP}} \left(X_i(t) - H_0 \right). \tag{5.4}$$

Synaptic scaling is introduced for homeostatic purposes and enforces that all incoming connection strengths are normalized to sum up to 1:

$$W_{ij}(t) = \frac{W_{ij}}{\sum_j W_{ij}(t)}. \tag{5.5}$$

As will be discussed later (Sect. 5.8.5 and 5.8.1), principles P4 and P5 emerge automatically from the architecture of the model.

Summary of reservoir behavior

Lazar et al. (2009) analyzed the behavior of the reservoir that will be summarized in the following. In contrast to our model, the reservoir was extended with readout neurons that were trained in a supervised fashion to predict the next input, i.e. a readout neuron was required to be activated if and only if the corresponding input neuron subset will be activated in the next time step.

- The network outperformed static reservoirs in a "counting" task. The network showed considerable increase in memory capacity and was able to predict a change in the input even after 20 time steps of constant input.
- Hierarchical clustering of the reservoir states showed the spreading of input conditions over many clusters of the reservoir states. For example, in the "counting" task, when presented an input sequence 'abbbb...bc' it is important to map the 5th and the 6th 'b' onto different reservoir states in order to be able to discriminate them. The network achieves that more effectively than a static reservoir that tends to lump several input conditions into the same clusters.
- A principal component analysis of the reservoir states revealed that the network learns to map each input condition into a tight cluster of reservoir states while the states move further apart from each other as learning proceeds thus enabling an effective representation of input conditions.
- As learning proceeds the network dynamics moves to a stable regime as revealed by perturbation analysis.
- All reservoir neurons become active at some point as opposed to random reservoirs that often contain a high percentage of inactive neurons.
- Reservoir neurons tend to become more input specific and tuned.
- Homeostatic mechanisms turn out to be crucial for healthy network dynamics. Without synaptic normalization the network develops seizure-like bursting activity where the neurons' activities are highly correlated. Without intrinsic plasticity neurons tend to become either hyperactive or not active at all.

For a more detailed computational modeling and theoretical analysis the reader is referred to Lazar et al. (2009). Subsequently, we will discuss our reinforcement learning extension of the reservoir.

Reinforcement learning architecture (RLA)

We extend the original network by Lazar et al. (2009) by replacing the readout neurons by action neurons and by implementing a version of on-line temporal difference learning. Basically, the reinforcement learning architecture takes the high-dimensional reservoir states reflecting the current input and gaze position and maps them to one of two possible actions (e.g. move gaze to left or right). Accordingly, the gaze position at the next time step corresponds to the action (eye movement) performed. As principle P1 demands, the looking time at a stimulus, i.e. the number of occurrences where the gaze equals the input position, is maximized by the reinforcement learning mechanisms.

More specifically, let n be the number of input types, e.g. $n = 2$ if an input can occur left or right. Accordingly, there are also n gaze positions and n possible actions (the action "no action" is modeled implicitly if no action unit is activated).

Each reservoir unit is connected to the action units by the connectivity matrix V. An action a is picked from a softmax probability distribution

$$p(a) = \frac{\exp(Q_a/T)}{\sum_b \exp(Q_b/T)}, \tag{5.6}$$

where the action value is given by $\vec{Q}(\vec{X}) = V \cdot \vec{X}$ and T is the softmax temperature. The action a determines the gaze position g at the next time step

$$g(t) \equiv a(t-1).$$

At each time step the network receives a reward defined by

$$r = \begin{cases} +1 & \text{if } i = g \\ -1 & \text{if } i \neq g \end{cases}$$

where i denotes the current input, e.g. $i = 1, 2$ for inputs "left" or "right". This reward and the action values of the current and next time steps are used to calculate the temporal difference error

$$\delta = r + \gamma \max_{a'} Q_{a'}(\vec{X}') - Q_a,$$

where $\vec{X}'$ denotes the next state of the reservoir and γ is the discount factor. The temporal difference error δ can then be used to globally modulate the weights V_{ak} from reservoir unit k to action unit a of the actor:

$$\Delta V_{ak} = \alpha \, \delta \, X_k,$$

with α being the learning rate. This algorithm is a simple form of on-line gradient-descent TD(0) learning with an action-value function linear in the states. The reader is referred to the standard literature for details (see, e.g. Sutton and Barto (1998), chapter 8).

Model parameters

The following table summarizes the model parameters.

Parameter name	Variable	Value
Number of neurons	N^E	60/80
Number of neurons per input	N^U	8
Connection probability between two excitatory neurons		0.2
Average firing rate	H_0	0.27/0.20
STDP learning rate	η_{STDP}	0.001
IP learning rate	η_{IP}	0.001
Discount factor	γ	0.7
Reinforcement learning rate	α	0.05
Softmax temperature	T	1.0

Note: 60/80 means that in all experiments the network consisted of 60 units except

Experiment 2 with 80 units since three different input and gaze populations were necessary and too few neurons would be left for the reservoir otherwise.

5.5 Experiment 1: Modelling the LR, LLR, LLLR and IR sequences

5.5.1 Modelling procedure

In Experiment 1 we train the network with one of four kinds of sequences: a left-right alternating sequence (LR), sequences where the left input is shown two or three times before the right input (LLR and LLLR) or a sequence made up from all of the mentioned sequences that are picked randomly with equal probability (e.g. LLRLRLRLLLRLRLLR...) which we call irregular sequence (IR) in accordance with Canfield and Haith (1991).[5]

During training the input neurons of the network receive additional drive according to the sequence that the network is trained to learn. The performance of the network is tested regularly for 100 time steps as learning proceeds. During this time the softmax temperature is reduced exponentially from $T = 1.0$ to $T = 0.1$ (simulated annealing) in order to reduce stochastic exploration of actions with time which is necessary for the system to reach optimal performance. During the last 50 time steps the gaze behavior of the network is recorded.

5.5.2 Results

Fig. 5.4 exemplifies the model's behavior after successful learning of the LLLR sequence. In this example, the time ranges from 1 to 30, as shown at the bottom of the figure. Subfigure (A) shows the input sequence given to the network: L-L-L-R-L-L-L-R-... Subfigure (B) shows the activity of the network reservoir consisting of

[5]Use of language: throughout the section the term "experiment" is meant to refer to modelling experiments as opposed to the term "study" which denotes infant experiments.

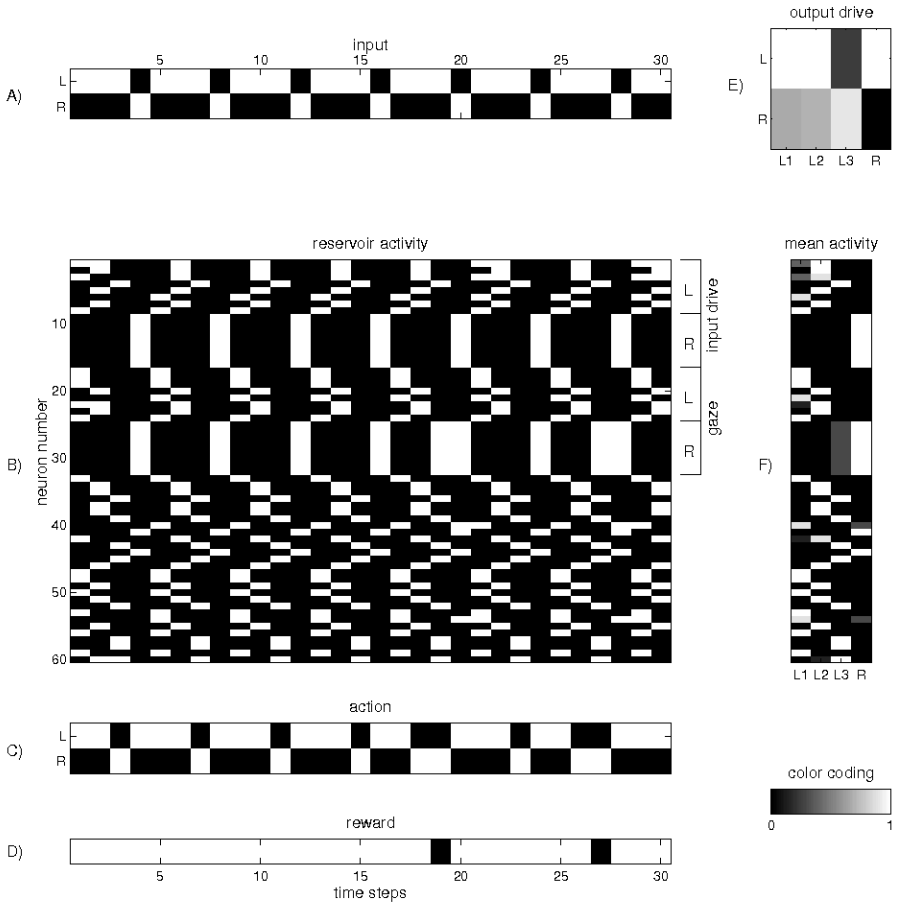

Figure 5.4: Network activity after successful learning. **A:** Input sequence LLLR. **B:** Reservoir activity. The first 16 units receive additional input drive according to the input sequence in A), the second 16 units receive additional gaze drive. **C:** Actions performed by the network. One time step after each action the gaze shifts to the corresponding position, as can be observed in B). **D:** Reward for the network. White: reward $= +1$; black: reward $= -1$. **E:** Output drive denoting the mean activity vectors (F) times the action unit matrix V. Reflects the net input for the output units. **F:** Mean activity vectors from B) for each input situation L_1, L_2, L_3 and R.

60 neurons as shown on the left side of the subplot. White and black dots denote active and non-active neurons, respectively. The first 16 neurons are input neurons that represent the LLLR input sequence, while the first 1-8 neurons tend to get

active when the input is L and the neurons 9-16 are activated when the input is R. Note, that not all of the neurons 1-8 get active because the intrinsic plasticity rule adjusts the activation theshold such that a predefined mean firing rate is sustained. In addition to the input sequence neurons, there are gaze neurons 17-32 that are activated according to the action performed a the previous time step (C). For example, at the *third* time step, the action R is performed (subplot C). Therefore, the gaze units 25-32 get activated at the *fourth* time step (subplot B). Analogously, when at the fourth time step, the action L is performed, (most) gaze units 17-24 are activated at the fifth time step. In this way actions influence the network activity. As a whole group, the units 1-32 drive the network. This activation spreads through the recurrent network and leads to a series of activity patterns that seem to converge to a cycle of four patterns (reservoir states) after learning (F). We see, that the network pattern repeats every fourth time step, which corresponds to the sequence length LLLR that repeats over and over. Averaging the activity of the time steps 1, 5, 9, 13, ..., 28 leads to the first pattern L_1 in subplot (F). Similarly, averaging over the time steps 2, 6, 10, ..., 29 defines the second pattern L_2, and so on. After learning the activity seems to have converged to these four cluster centers and cycles through them indefinitely (see Sect. 5.8 for further investigation of this issue).

The actions are selected through matrix V which maps the reservoir activity to the action units (e.g. move gaze to left or right), Fig. 5.3. Subsequently, this output drive (E) is used to select the actions (C) by the softmax rule (5.6) (Fig. 5.4).

Since the matrix V is initialized randomly, the model picks random actions in the beginning and therefore performs also random gaze shifts. This is not the case any more after learning in Fig 5.4. Subplot (B) shows that the gaze unit activities (17-32) correlate strongly with the input unit activities (1-16). This is exactly what the reinforcement learning architecture accomplishes: it adjusts the action connection matrix such that the correct gaze shifts are performed in order to match the input unit activity. Such behavior is rewarded as depicted in subplot (D) where the obtained reward is almost always +1 (white).

In the following we define several measures based on the network's gaze behavior. The results of this behavior in the course of learning are shown in Fig. 5.5.

Performance

The performance is measured as the average reward received by the network during the test phase. Since the reward can be -1 or $+1$ the performance also ranges between -1 and $+1$.

Reaction time

Since we do not use inter-stimulus intervals and each stimulus lasts for one time step the network can show only two types of behavior towards a stimulus: anticipatory and reactive. They refer to the network initiating an eye movement one time step before or concurrently with the occurrence of the right stimulus, respectively. Accordingly, anticipatory eye movements are assigned a reaction time of -1 and reactive eye movements $+1$.[6]

Percent anticipation

This variable measures the proportion of anticipatory eye movements in relation to all eye movements to the right stimulus (R).

[6]These values are chosen such that a random eye movement policy leads to zero average reaction time.

Gaze shifts to the right/correct stimulus

These measures denote the proportions of gaze shifts to the right/correct stimulus. For example, in the LLLR sequence it is measured how often the gaze is shifted to the right/correct stimulus after the occurrence of the first left, second left, third left or right stimulus. "Correct" refers to the cases when the anticipatory gaze shifts are followed by the actual occurrence of a stimulus at the anticipated position.

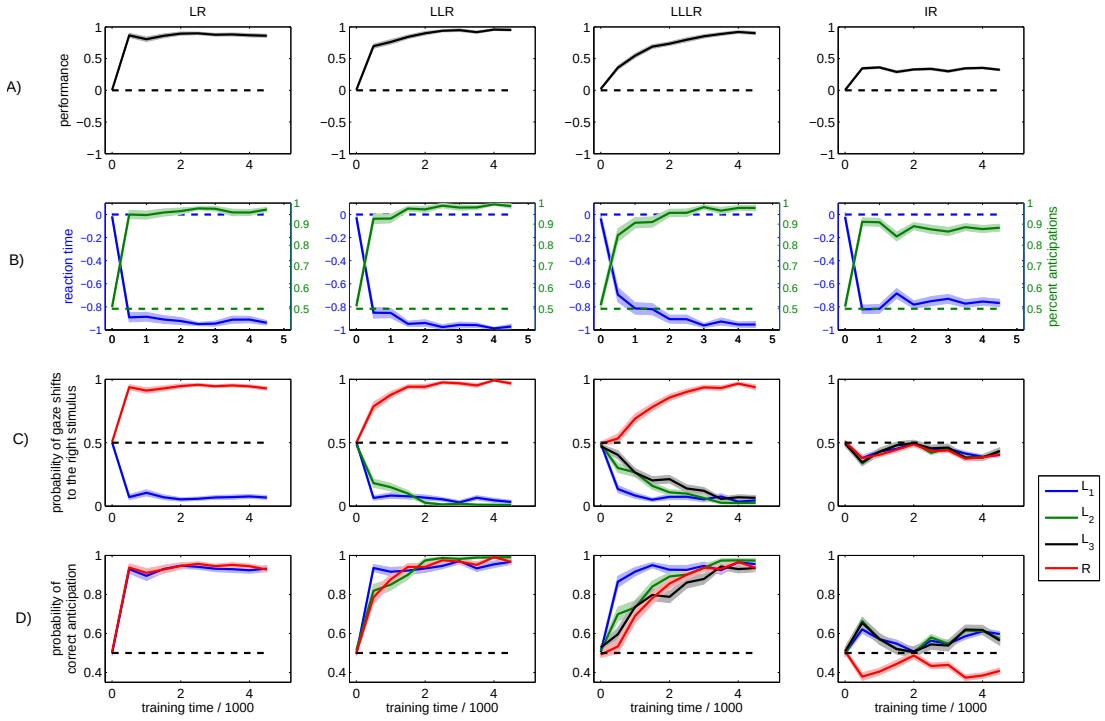

Figure 5.5: Results of Experiment 1: LR, LLR, LLLR and IR sequences as a function of training time. Dashed lines denote the chance level. Shaded areas denote the standard errors after 100 simulations. **A:** Average reward. **B:** Reaction time and percent anticipations of the right stimulus. **C:** Probability of a gaze shift to the right while the actual input is the first left (L_1, blue), second left (L_2, green), third left (L_3, black) or right (R, red) stimulus. **D:** Probability of a correctly anticipated gaze shift to the first left (L_1), second left (L_2), third left (L_3) or right (R) stimulus.

The results in Fig. 5.5 can be summarized in the following way.

1. For all regular sequences, with increasing training time,
   a) the reaction time decreases and saturates at some point (Fig. 5.5B),
   b) the probability of correct anticipations increases while the probability of wrong anticipations decreases (see Fig. 5.5B (percent anticipations), C and D).
2. As sequences get more complex
   a) the reaction time increases (Fig. 5.5B),
   b) the differences between reaction times toward sequences of different complexity first increase with training time and then decrease again as training time grows large (Fig. 5.5B),
   c) the probability of correct anticipatory gaze shifts to the right decreases (see Fig. 5.5C, red curves, and D).
3. The probability of an anticipatory gaze shift to the right increases with each presentation of a left stimulus (see Fig. 5.5C).
4. This result will be mentioned below in Experiment 2.
5. The probability of correct anticipations decrease from L_1, L_2, L_3 to R (Fig. 5.5D).

5.5.3 Discussion

The first three results conform to experimental observations summarized in Sect. 5.2. Result 1) can be interpreted in two ways: the training time can be seen as modelling both the age of the infant and the number of trials presented to the infant during a session. Both interpretations are supported by experimental data. On the one hand, the most natural interpretation of training time is the trial number since infants get repeatedly exposed to the sequences during the

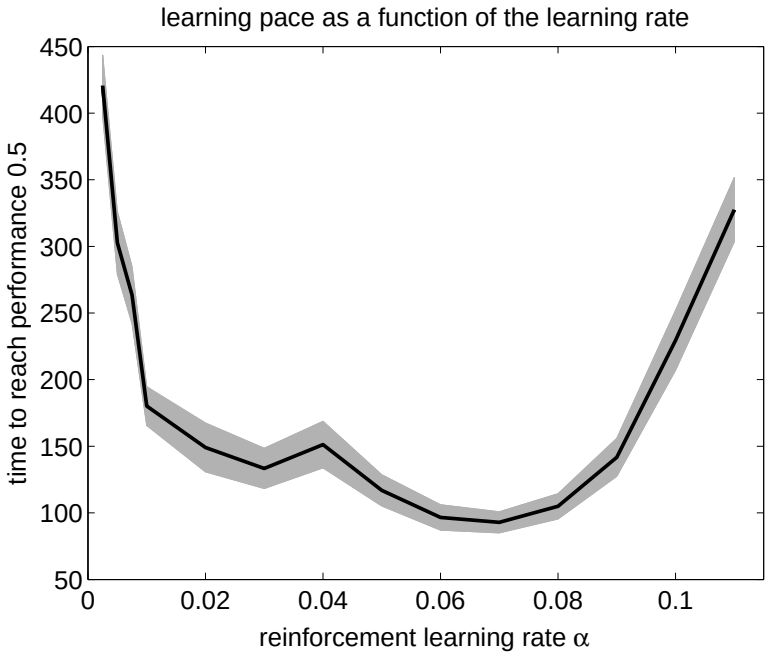

Figure 5.6: Learning pace as a function of the reinforcement learning rate α. The y-axis shows the average point in time when the model reaches performance 0.5. Data was acquired for the LR sequence and averaged after 100 stimulations.

study. On the other hand, the age of the infant can be modeled as learning rate of the algorithms. Indeed, infants' reaction time decreases with age even for unpredictable sequences which has been suggested to due to increasing processing speed (Canfield et al., 1997). Fig. 5.6 shows that, as expected, the LR sequence as acquired faster as the learning rate increases up to values of $\alpha = 0.07$ (after that learning can not converge because the learning rate is too high). Therefore, for small learning rates, the results in Fig. 5.5 would be stretched, i.e. the network needs more time to reach the same performance levels. In this way, varying learning rates account for age differences. Therefore, as we are pursuing a qualitative account, training time will be considered as modeling both trial number and infant age in the following.

Analogously to infants, the model develops spatiotemporal expectations by coping with delayed reward and initiating anticipatory gaze shifts before the onset of stimuli (compare Fig. 5.4B and C). This leads to an increase of the percent-

age of anticipations (1b) and a necessary corresponding decrease in reaction time (1a) since any anticipation counts as reaction time of -1. We see this observation confirmed by Fig. 5.5B where the increase in anticipations mirrors the decrease in reaction time. Note that the irregular sequence is not without structure since it is made up of randomly picked LR, LLR and LLLR sequences which carries a lot of probabilistic information.

As for result 2a), the proposed model behaves similarly to infants since learning the structure of a longer sequence requires more memory (the various L's need to be distinguished from each other) and therefore more learning time. Note, that thereby the model automatically captures principle P4 without the need to build it in purposefully (see Sect. 5.8.5 for details). The age differences (2b) are also explained by the model in the following way. For all sequences the gaze behavior is random at the beginning of the training. Therefore, there has to be a gradual increase in the differences between the various sequence types in the course of training. As training time grows large though, the differencies diminish again since the reaction times saturate at the same level for all deterministic sequences. This constitutes a prediction of the model. Result 2c) mirrors result 2a) and is therefore not surprising given that the reaction time increases with the complexity of the sequences.

Since principle P4 is implemented by the model we have already explained how it accounts for result 3) in Sect. 5.3. Result 5) is a direct consequence of this principle and constitutes a prediction of the model. In Sect. 5.8.5 it will be explained how the computational model realizes this principle.

5.6 Experiment 2: Modelling the pivot sequence: left-top-left-bottom

In order to gain a more general view on the development of visual expectations it is instructive to look at sequences other than just left and right alternating ones.

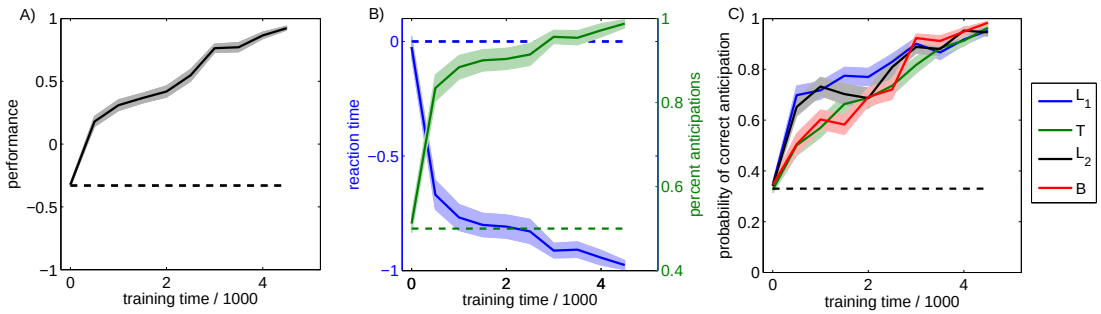

Figure 5.7: Results of Experiment 2. Dashed lines denote the chance level.

5.6.1 Modelling procedure

Following an experimental study by Reznick et al. (2000) the stimuli can occur at three possible positions: left (L), top (T) and bottom (B). The stimuli are displayed in a deterministic manner: L-T-L-B-L-T-L-B.... Accordingly, the model is endowed with three position inputs, three gaze inputs and three gaze shift outputs with eight reservoir units reserved for each position resulting in 48 input units in total. The network size was increased to 80 units. Otherwise all parameters were kept at the same value as in Experiment 1.

5.6.2 Results

All experimentally observed results are confirmed by our model:

1. Reaction time decreases with training time while the percentage of anticipations increases (Fig. 5.7B).

4. The probability of correct anticipations is higher for the pivot stimulus (left) than for either of the wing stimuli (top and bottom) (Fig. 5.7C).

5.6.3 Discussion

As discussed in the previous section result 1) is a consequence of successful learning rewarded when the gaze is a priori shifted to the correct stimulus position.

Result 4) confirms directly the corresponding experimental result. We observe that for the pivot sequence as well as the previous sequences the model behaves according to principle P4: since the pivot stimulus occurs twice as often as either of the wing stimuli and learning the occurrence of the pivot stimulus is easier because it requires less memory.

5.7 Experiment 3: Learning and relearning

Facilitated reacquisition of already learned tasks has been observed in rabbit conditioning (Frey and Ross, 1968) and infant memory retention tasks (Rovee-Collier, 1999). Infant memory is usually studied with the mobile task where infants learn to move a crib mobile by kicking a ribbon strung between a mobile hook and one ankle. Retention of memory is measured via an increase in the average kick rate.[7] As for visual expectations the question whether infants learn the structure of a sequence faster if they already learned it in the past as compared to learning it for the first time has not been investigated yet. The following experiment uses our model to predict the result of such a future study which can serve as a test for the model.

5.7.1 Modelling procedure

The network is exposed to the LR sequence for 1000 time steps which is enough for the performance to saturate. Subsequently, the input sequence is switched to

[7] This task can also be viewed as an instrumental conditioning task although there is debate about the role of implicit vs. explicit memory in the definition of the task (Rovee-Collier, 1997).

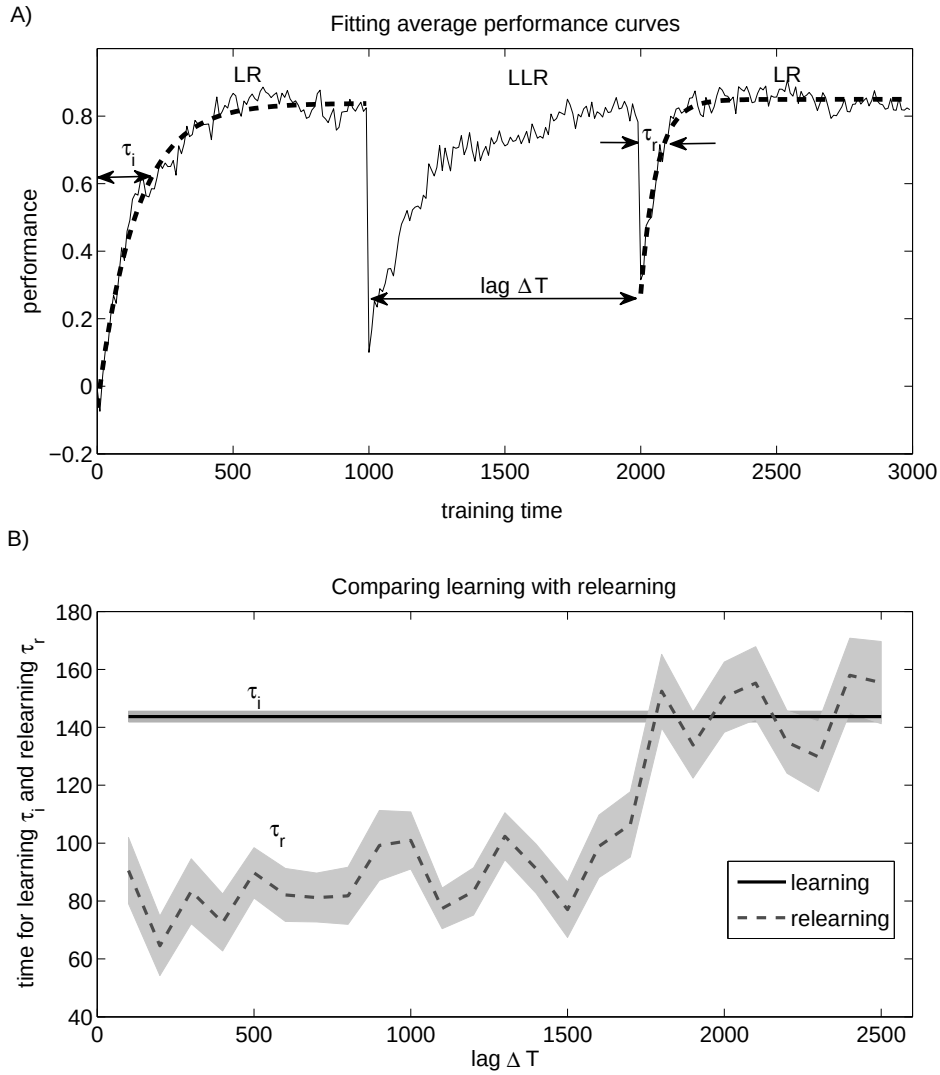

Figure 5.8: Learning and relearning of the LR sequence. **A:** Average performance curves from 100 simulations fitted with a least squares fit of the function $f_{a,b,\tau}(t) = a\ \exp(-t/\tau) + b$. **B:** Widths τ of the exponential fits of learning the LR sequence for the first time versus relearning it again after ΔT time steps. Gray shaded areas denote the standard errors.

LLR for ΔT time steps after which the LR sequence is switched on again.

5.7.2 Results

Fig. 5.8A shows the mean performance curve of the model based on 100 simulations. Each time when the input sequence is changed there is a drop in the performance. Even by visual inspection it can be observed that the increase in performance is faster for relearning than for learning. This result is confirmed quantitatively by measuring the widths of fitted exponential curves to the performance data (Fig. 5.8B), where the relearning time stays significantly below the learning times for a wide range of lags. Even after 11-12 times the average learning time (145) has passed, relearning is still faster ($1700 \approx 12 \times 145$).

5.7.3 Discussion

It is not easy to explain this behavior of the model. We conjecture that it is due to the characteristics of the reinforcement learning architecture: it takes time to build and rebuild the mapping from the reservoir to the action units, i.e. the action connection matrix V needs time in the order of 1500 time steps to change (Fig. 5.8B)). Also, it is difficult to destroy a working network configuration (LR sequence) since it requires to solve a hen-egg problem: the new reservoir states necessary for learning the action connection matrix are difficult to learn since a frequent wrong gaze unit input mediated by a yet wrong action connection matrix disturbs the formation of the new reservoir states. Consequently, a once learned representation is upheld for a long time as observed in Fig. 5.8B.

5.8 Analysis of model behavior

In the previous sections we successfully modeled the LR, LLR, LLLR, IR and the pivot sequences. In the following the model's behavior will be studied in more detail in the case of the LLLR sequence. The analysis consists the following steps. In Sect. 5.8.1 it will be shown that the reservoir activity develops accoring to the

main results of Lazar et al. (2009) described in Sect. 5.4.2. Specifically, it will be shown that the reservoir states cycle between four cluster centers after learning. In the subsequent section the construction of the reinforcement learning architecture (RLA) will be discussed and justified. In Sect. 5.8.3 the coupling between the reservoir and the RLA will be analyzed through the study of random gaze positions. The subsequent section will show that the reservoir state clusters that develop after learning are actually suitable for reinforcement learning. Finally, in Sect. 5.8.5 we will investigate how the model realizes the important principle P4 and relate it to various obtained results.

5.8.1 Development of the reservoir activity

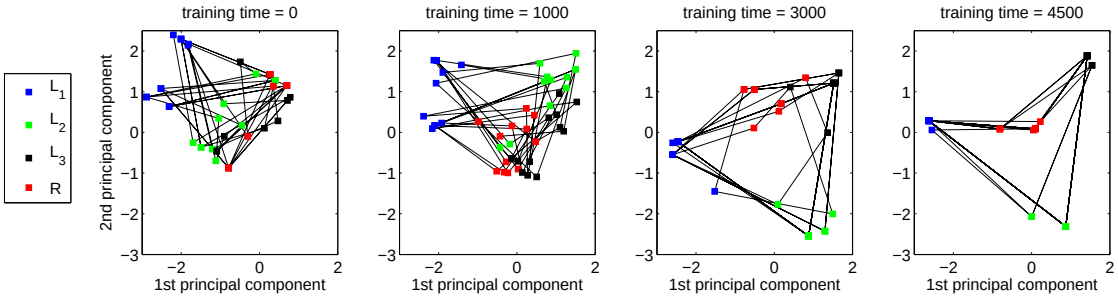

Figure 5.9: Principal components of the network activity in the LLLR sequence. The first two principal components account for 60% - 70% of the variance.

Fig. 5.4B) and F) suggested that the reservoir activity cycles through four clusters of states (F) in the reservoir state space. The development of this behavior is shown in Fig. 5.9 that visualizes the projection of the reservoir activity onto its first two principal components[8]. Each point in the figure corresponds to a state in the reservoir state space. At the beginning of the training (training time = 0) the reservoir responds first unstably to the input sequence LLLR. The reservoir activity moves irregularly through the state space. Unsupervised learning with

[8]Principal Components Analysis (PCA) is a mathematical technique for mapping high-dimensional data such as the 60 dimensional reservoir state to a low-dimensional space that captures the highest variance of the data. For example, this enables plotting high-dimensional data on the two-dimensional plane, as shown in Fig. 5.9.

STDP makes the reservoir states that correspond to the same input role (L_1, L_2, L_3, R) converge to distinct cluster centers, as was discussed in context of Fig. 5.4F. This is also reflected by Fig. 5.9, training time = 4500, where the reservoir activity visits basically only four distinct states. These cluster centers move further apart as learning proceeds which means that the clusters in Fig. 5.4F become more and more distinct with time. This corresponds to the requirement of principle P5: the reservoir patterns that correspond to the various stimuli move apart from each other as learning proceeds. Once the pattern sequence is learned the network is autonomously able to sustain its activity sequence even in the absence of inputs thereby predicting future input states. Note that although during L_1, L_2 and L_3 the input to the network is the same, the network manages to treat them differently, thereby implicitly "enumerating" them. This behavior is not trivial since the network treats the various presentations of a left input differently (L_1, L_2 and L_3) while treating the corresponding inputs at the subsequent trials similarly (e.g. at each trial L_1 corresponds to the same cluster center). This is due to the fact that during the presentations of L the reservoir's recurrent activity assigns different states to the L's while the presentation of R strongly perturbs this activity basically resetting the reservoir state back to L_1 (see also discussion after the introduction of principle P4 in Sect. 5.3).

5.8.2 Construction of the reinforcement learning architecture (RLA)

The extension of the reservoir by the RLA is not trivial because of the RLA's feedback influence on the reservoir. As was observed in our simulations, even in absence of the RLA the reservoir develops four distinct state clusters in the case of the LLLR sequence which is in accordance with previous results (Lazar et al., 2009). In a simplified construction without gaze units it has been tried to map the four points in the reservoir space to action units using a reinforcement learning procedure that learns the weights of the connection matrix V. Although this feedforward extension would not have complicated the model because of the absence of feedback to the reservoir, a closer look reveals that proper learning would not

be possible for the following reason. The model is based on our hypothesis that infants receive a reward whenever their gaze position equals the input position. Therefore, in order to define the reward the gaze position has to be represented in the reinforcement learning state space.

As a second trial, one could consider to represent the gaze position by a variable separate from the reservoir activity. Roughly speaking, the reinforcement learning state space will now consist of eight states, two gaze states times four reservoir states. In that case the gaze position could be used in order to define the reward but it would not be reflected by the reservoir activity in any way. This bears the problem though that learning will not converge since it requires eight *distinct*, mostly non-overlapping states. In our case though the reservoir activity is the same for both gaze positions. Since the gaze position determines the reward, the *same* connection weights of the matrix V to this reservoir activity will sometimes be strengthened and sometimes weakened depending on the reward. Therefore, proper learning and convergence is hindered by this construction. Fig. 5.11C) shows how the performance of reinforcement learning in a simplified model (see below) declines with increasing overlap of the states. Since the reservoir activity is the same for either gaze position, the maximal overlap (defined below) is 1, therefore learning must fail as witnessed by Fig. 5.11C).

For these reasons it is necessary to acquire eight distinct reservoir states for each gaze $\times$ input combination. Thus, it seems necessary to include the gaze units into the reservoir that are determined by the RLA's actions.

5.8.3 Coupling the RLA to the reservoir

In order to investigate the roles of the reservoir separately from the RLA and the role of their coupling, the feedback from the RLA to the reservoir was decoupled for the first 5000 time steps during which the gaze position was picked randomly with uniform probability. After 5000 time steps the RLA was coupled again to the reservoir, i.e. the action unit activities determined the gaze positions at the

subsequent time step.

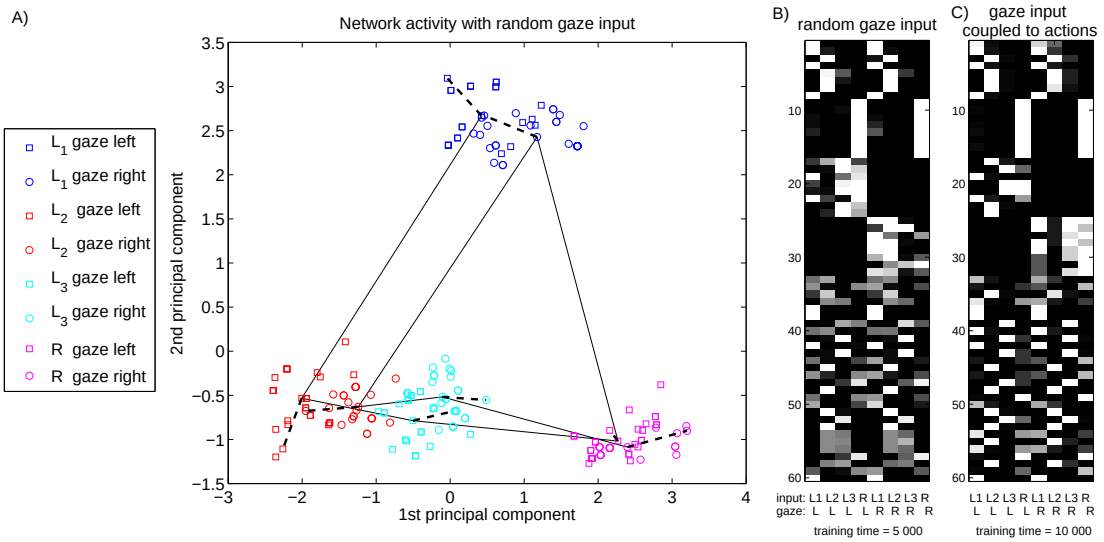

Figure 5.10: **A:** Principal components of the network activity in the LLLR sequence with random gaze positions. The first two principal components account for 60% - 70% of the variance. The states were grouped according to the current gaze × input combination. Dashed lines denote the change of cluster centers after 5000 steps of learning after coupling the RLA to the gaze units. **B and C:** Cluster centers of reservoir activity before coupling and 5000 time steps after coupling.

Fig. 5.10A) shows the principal components of the reservoir activity from time steps 4801 until 5000, where the states were colored according to the current input × gaze combination. Indeed, the reservoir activity has fallen into eight clusters even though the gaze positions were picked randomly. The figure also shows that even after coupling the gaze units back to the RLA the eight cluster centers do not change their position much as depicted by the dashed lines although the RLA quickly reaches a high performance level after coupling (Fig. 5.11A). This is confirmed by direct observation of the cluster centers in Fig. 5.10 before (B) and after (C) coupling and learning for another 5000 time steps. We conclude that the reservoir is able to develop eight distinct clusters both during the random gaze phase and keep the same states during learning with the gaze positions coupled to the RLA.

This situation seems to simplify the task for the RLA considerably since the coupling is not dynamically changing which would have required the RLA to adapt constantly. In the contrary, the task for the RLA is thus reduced to learn the appropriate mapping from the eight mostly static reservoir states to the actions. In the following it will be shown that the RLA is able to succeed in this task.

5.8.4 Relation between learning performance and the state overlap

The introduction of gaze units into the reservoir was motivated by the requirement of eight distinct states for reinforcement learning to be possible. In this subsection we investigate whether the reservoir states as depicted in Fig. 5.10B) and C) are actually distinct enough.

Let $\{\vec{X}_1, \vec{X}_2, ..., \vec{X}_n\}$ be a set of N^E-dimensional vectors. Then we define the overlap between two vectors i and j as the cosine of the angle between those vectors:

$$\Omega_{ij} \equiv \frac{\vec{X}_i^T \cdot \vec{X}_j}{||\vec{X}_i|| \cdot ||\vec{X}_j||}. \tag{5.7}$$

The overlap Ω of the whole set of vectors will be defined as

$$\Omega \equiv \max_{i,j} \Omega_{ij}. \tag{5.8}$$

Fig. 5.11B) shows the development of the overlap of the eight cluster centers (compare Fig. 5.10B and C) as a function of training time. Especially before the coupling of the gaze units to the RLA at 5000 time steps the maximal overlap ranges at around 0.9 begging the question of whether the reservoir states are distinct enough for learning.

In order to investigate that question we implemented a separate model using the same reinforcement learning architecture but without the reservoir. The eight

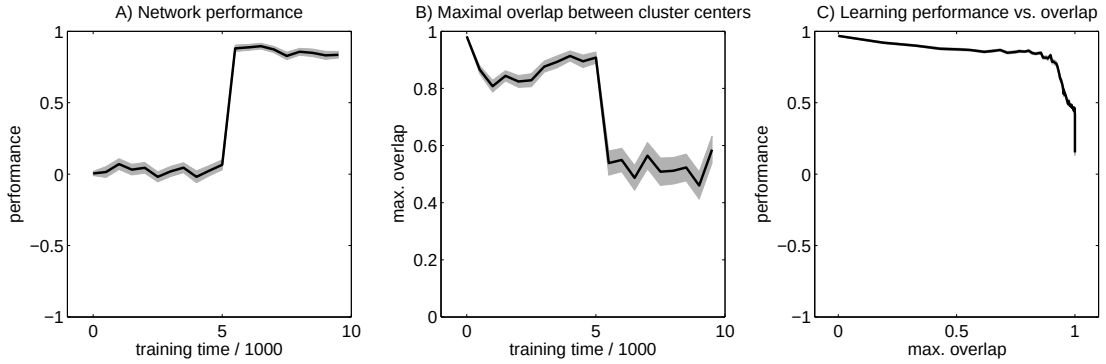

Figure 5.11: Relation between reinforcement learning and reservoir state overlaps. **A:** Reinforcement learning performance as a function of training time. The gaze states are coupled to the RLA after 5000 time steps. **B:** Maximal overlap Ω as defined in eq. (5.8). **C:** Relation between learning performance and the overlap Ω obtained by a separate simulation.

$N^E = 60$-dimensional states were created artificially according to the following procedure. The first $\omega \in \{0, 1, ..., N^E\}$ components of each vector was set to 1. For each remaining component one and only one vector component was set to 1 resulting the following exemplary scheme:

$$\begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ \vdots \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ \vdots \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ \vdots \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ \vdots \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \end{pmatrix}. \tag{5.9}$$

We run 100 simulations for each value of ω and calculated the average overlap of the vectors. The reinforcement learning performance was measured as the average reward during the last 50 time steps of the total 300 time steps of the simulation. Plotting the overlap against the performance resulted in Fig. 5.11C).

We observe that even for overlaps as high as 0.9 the reinforcement learning per-

formance is above 0.8 (corresponding to 90% because of the scaling between -1 and +1) and declines quickly as the overlap approaches 1.0. In this simulation the state vectors reflected the "worst case" since they we required to have a pairwise overlap of Ω instead of only some pair of vectors. We conclude that even in the worst case scenario, an overlap of 0.9 is low enough to reinforcement learning to succeed.

In summary, our analysis of the coupling between the reservoir and the RLA leads to the following picture. The reservoir captures the structure of the sequence inputs by mapping them onto eight reservoir states — four for each of the LLLR input and two for each of the gaze positions even though at the beginning the gaze position may fluctuate randomly due to the initialization of the RLA. The eight reservoir states remain sufficiently stable during learning therefore enabling the RLA to learn the appropriate mapping from the reservoir states to the action units. A separate simulation showed that the RLA can cope with this problem even up to state overlaps of 0.9. The analysis of our model showed that the actual overlaps are low enough for the RLA to get started with learning. As Fig. 5.11B) shows, the overlap decreases significantly to values around 0.5 as learning proceeds and thereby facilitating learning even more. The decrease of the overlap reflects that the model learns to visit only the four (less overlapping) rewarding states out of the eight states in agreement with Fig. 5.9.

5.8.5 Development of correct anticipations

At this point, it is useful to discuss the development of correct anticipations and their ordering in Experiments 1 and 2. In Figs. 5.5D and 5.7C the ordering of correct anticipations depends heavily on the sequence and the position of the stimulus in agreement with experimental results (Canfield and Haith, 1991, Canfield and Smith, 1996, Reznick et al., 2000). Specifically, the probability of correct anticipations does not develop equally fast for different sequences and positions in the sequence, e.g. the anticipation of L_1 develops faster than L_3 (Fig. 5.5D, LLLR). In Sect. 5.3 it has been shown that principle P4 can account for these differences

demanding that the prediction ability of a stimulus be dependent on the required memory necessary to unambiguously determine it.

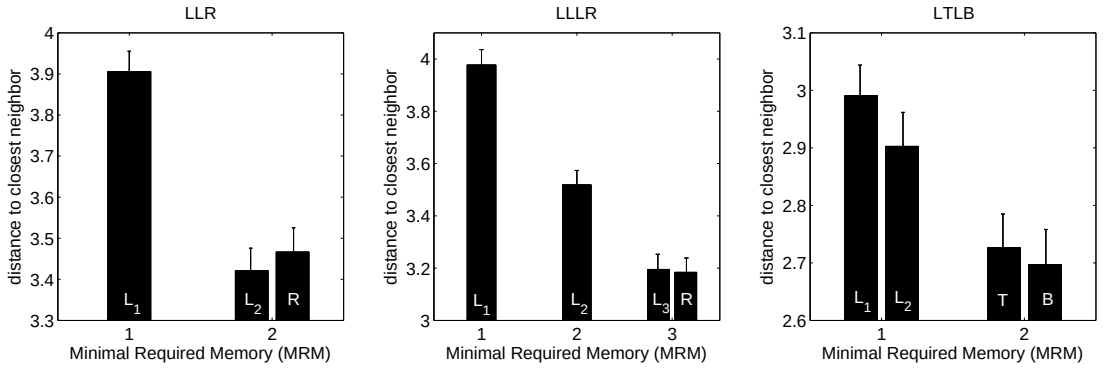

Figure 5.12: Euclidean distances of the activity cluster centers to their closest neighbors. For each sequences the network was run 100 times for 2500 time steps and the mean activity after learning was measured and grouped according to the input stimuli. The MRM of each sequence element was determined according to the definition of MRM in Sect. 5.3.

We quantified the behavior of our model in Fig. 5.12 where the separation of a cluster center from all the others was measured by calculating the Euclidean distance of each cluster center to its closest neighbor. The figure shows that for all sequences the distances decrease as the MRM of a stimulus increases in agreement with principle P4. Note that the network was run for only 2500 time steps where only intermediate performance is achieved (compare Figs. 5.5 and 5.7). This is due to the observation that after learning has converged, the cluster centers move farther apart from each other and the differences in their distances fade away which is in agreement with principle P5.

In order to investigate the relation between the distances of the cluster centers and the MRM, we studied the behavior of the reservoir further. As quantified above, stimuli that are uniquely determined by the previous stimulus are assigned a state far away from the other states, since all other states are driven by varying preceding input stimuli by assumption. Therefore, stimuli requiring less memory can also be anticipated more easily, e.g. the pivot stimulus L in the LTLB sequence.

On the other hand, if a stimulus requires more memory (e.g. T), i.e. the previous stimulus (L) does *not* determine it completely, it will be mapped to a similar state in the state space as all the other stimuli that can also be successors of L, e.g. B. Therefore, T and B can not be distinguished from each other as easily as the state L from T and B which leads T and B to be anticipated less successfully than L as reflected by Fig. 5.7C. Taken together, this behavior of the model realizes principle P4.

5.9 General discussion

In this chapter we reviewed experimental data on visual expectations in infancy and developed a theory of visual expectations based on five principles. We implemented a computational model according to those principles. The model successfully explains infants' reaction time and anticipation behavior during the left-right (LR) alternating sequence, the asymmetric sequences LLR and LLLR and the more complex pivot sequence (left-top-left-bottom). Through a synergistic combination of plasticity mechanisms the network maps the sequence of inputs to distinct reservoir states and thereby enumerates repeating states. A reinforcement learning architecture learns to perform accurate eye movements that enable the comparison to infant data. It can also cope with the delayed reward given for a successful fixation of the stimulus which is necessary for anticipatory eye movements.

5.9.1 Related work

Our model is novel from several points of view. First, to our knowledge there are no other computational models for the development of sequence learning in infants in the literature although models of other forms of visual expectations like predictive gaze control for object trajectories (Balkenius and Johansson, 2007) do exist.

Second, as far as we know the model is the first fully recurrent neural network used in developmental psychology of infants. Elman networks (Elman, 1990) have been used widely including our work on causality and occlusion perception (Chapter 2) but they are only partly recurrent. Elman networks (Lin and Mitchell, 1992) and variants (Saeb et al., 2009) using reinforcement learning have also been suggested but not in the context of infant research.

Third, the model is also novel from the technical point of view. Artificial neural networks have widely been used as function approximators in reinforcement learning for maintaining the value function of an agent (Tesauro and Sejnowski, 1989, Bertsekas and Tsitsiklis, 1996). On the contrary, only limited work has already been done using recurrent neural networks, probably because of difficulties in training such networks. Perhaps the closest work to ours has been done on reinforcement learning with echo state networks (ESN) (Szita et al., 2006). Contrary to ESNs or liquid state machines (Maass et al., 2002) that use a dynamic reservoir with fixed connection weights our model uses different interacting plasticity mechanisms and forms more effective representations than networks with fixed weights (Lazar et al., 2009).

Finally, our model differs from the model by Lazar et al. (2009) by using a reinforcement learning architecture to model the infants' actions (eye movements). Note, that the architecture is not only built "on top" of the reservoir but also influences the reservoir activity though the action dependent activity of the gaze neurons leading to non-trivial dynamics.

5.9.2 General account for sequence learning in infancy

The account of the model for sequence learning in infancy can be summarized in the following way. When infants are exposed to sequences of interesting visual stimuli they try to maximize the looking time at each of the stimuli (principle P1). We suggest that this principle guides the looking behavior only as long as infants are interested in the stimuli and the task. As observed by Haith and McCarty (1990), p. 73, infants' interest level tends "to reach a peak performance level, and

then their performance declines". Of course, on average infants will habituate to the sequence at some point. Furthermore, there are whole research areas on intrinsic motivation, saliency maps, familiarity vs. novelty preferences in infants etc. that certainly influence infants' gaze behavior. Principle P1 would have to be extended in order to incorporate these ideas but we choose to keep it at that for the sake of simplicity and for the purposes of our model.

As already discussed above, principle P1 leads to the necessity of predicting the stimulus sequence. In the context of our model, infants succeed in the prediction task in the following way. When exposed to two consecutive stimuli L and R two neuronal populations S_1 and S_2 emit spikes in an infant's brain (principle P2). Since S_2 follows S_1 in time the spike timing dependent plasticity (STDP) rule strengthens the synapses from S_1 to S_2 (principle P3). In future, when the infant is repeatedly exposed to stimulus L, the activity of neuronal population S_1 spreads via the strengthened synapses and activates population S_2 even without being driven by stimulus R. In this sense, it is possible to speak of prediction: the infant's brain has learned to predict stimulus R. Even if the same stimulus L is displayed repeatedly the network maps it to distinct activity patterns L_1, L_2 and L_3 (see Fig. 5.4B and F) because the activity of the recurrent network is constantly changing and evolving.

On top of this prediction reservoir a motor system reads the pattern sequences in such a way that appropriate gaze shifts can be performed to maximize the reward. Specifically, correct anticipations increase with time paralleling the network's ability to predict future stimuli. Similarly, the reaction time decreases due to increased anticipations.

5.9.3 Predictions of the model

1. The correct anticipations for any sequences are ordered by the principle P4 (see also Sect. 5.8.5). For example, the complex LRLLR sequence should be anticipated in the following way. Disambiguating it with $L_1R_1L_2L_3R_2$ it follows that percentages of correct anticipations should be ordered as L_3, $R_1 < R_2 < L_1$, L_2 since MRM(L_3), MRM(R_1) > MRM(R_2) > MRM(L_1), MRM(L_2). In other words, it should be easiest for infants (and adults) to

anticipate L_1 and L_2 and hardest to anticipate L_3 and R_1 while anticipating R_2 should be of intermediate difficulty.

2. Relearning happens faster than learning (Experiment 3).

3. Infants' looking behavior is independent of the particular stimulus position given the same sequence, e.g. a left-right alternation should not lead to different behavior than a up-down alternation (as observed by Reznick et al. (2000), Experiment 2).

4. As observed in experimental result 2a), the differences in reaction time to sequences of various complexity increases with age (e.g. from 2- to 3-month-olds). Our model predicts that these differences will decrease again as infants grow older (see Fig. 5.5B and corresponding discussion of result 2a)). We estimate that this should happen by the end of the first year of life.

5.9.4 Limitations and future work

One of the problems limiting a detailed modeling of infant studies is the difficulty of extending the model to realistic time scales. The durations used in experimental studies are usually of the order of 1000 ms. On the other hand the time scale of STDP is around 20 ms which has to be taken as corresponding to one time step. Therefore, a realistic model would use stimulus durations of roughly 50 time steps. Our model would not work with such long sequences because the network's memory is not large enough. Enlarging the network to more units may help but other than that scaling network performances to realistic scales is a general problem in neural network research which has to be tackled in future work. As for the trial numbers, the model reaches performance 0.5 (i.e. 75% correct actions) at around 90-100 time steps for the LR sequence (see Fig. 5.6) which corresponds to 45-50 trials. This is more realistic since it is off from real trial numbers by a factor of 1-3 only. Given that realistic anticipation rates are around 10-20%, i.e. far lower than 75%, our model works with realistic trial numbers.

Another limitation concerns the type of input representation chosen in the model. If the input units are to represent the human retina then eye-centred coordinates

would be more appropriate than the screen-centered ones chosen here. Specifically, the position of the input on the retina should be dependent on the gaze position. Successful learning would actually lead the input to stimulate the same space on the retina while the gaze direction would change according to the input sequence. This would complicate the model significantly, though, by requiring a transformation from eye-centered to head-centered coordinates. It is unclear if any reasonable lessons about the development of visual expectations may be learned from this effort.

In the context of real infant behavior it is important to ask about reactive saccades. Even before any prediction about the sequence can be made, infants show reactions to occurring stimuli by directing the gaze towards them. Specifically, their gaze behavior is far from random even in the beginning of the study in contrast to the model's behavior. In the context of the model it would be necessary to use a sequence with stimuli displayed for at least two times steps and to make the reinforcement learning architecture function faster than the reservoir learning mechanisms. In that way, maximizing looking time in absence of the ability to predict the sequence would lead to reactive saccades. This topic is indeed interesting and may be combined with the extension of the model to larger time scales in future.

Towards a model of enumeration

In Sect. 5.3 result 3) - the increase in probability to anticipate R after each presentation of L - was explained parsimoniously by principle P4 already accounting for other results. The explanation given was that for each L in the LLLR sequence the minimal required memory (MRM) grows thereby making its prediction less probable while making the opposite prediction (R) more likely. Note, that this account does not require any enumeration model suggested in the literature. Particularly, it is not a "number-based" explanation as put forward by Canfield and Smith (1996). We believe that our model can suggest an alternative view of enumeration in infancy and are excited to investigate this possibility in future work.

Chapter 6

Discussion and outlook

This thesis was motivated by the following questions. How does infants' cognition get started? What are the building blocks of initial knowledge and how do they develop? We decided to stay close to experimental data gathered in infant psychology, tried to sketch what a theory of infant cognition may look like and built computational models to test the plausibility of our ideas. We started by the parsimonious assumption that all infants have at birth is the "blooming, buzzing confusion" of light coming into the eye, the ability to move the eyes and some powerful learning mechanisms. Throughout the thesis we adhered to these assumptions while trying to conceive of how infant's first knowledge may come about.

The three projects in this thesis covered in total a broad set of data from around 25 studies that used various experimental paradigms like habituation, visual preference and visual expectation. They capture different phenomena ranging from the perception of causality, occlusion, object permanence, object tracking behind occluders, object unity and sequence learning. Despite this diversity all three projects embodied a common principle: the prediction of future events.

This principle suggests that what infants mainly do during cognitive development is building internal models and representations of the world and using these models to predict how the world will behave. The errors that they make during this process directly capture their attention and lead to further refinement of the models. Interestingly, this procedure essentially reflects the structure of scientific work.

The view that babies investigate the world basically in a scientific way is, of course, not new and has been put forward by the developmental researcher Alison Gopnik (see e.g. Gopnik et al. (1999)) as a research hypothesis on infant development. In summary, the guiding principle of prediction and anticipation of future events turned out to be very fruitful for the explanation of a broad range of data.

The first project in chapter 2 modeled the development of causality and occlusion perception since understanding that some objects move in a self-propelled way while others are caused to move, looked like a promising path to get started with the animate/inanimate distinction. This, in turn, could lead to the starting point of global categories that the world is divided into. We did not arrive at the latter goal since it seemed more important to understand even more basic principles like object unity more deeply.

The generality of the model formulation based on the prediction of future inputs made us conjecture that the general framework could allow for modelling other event categories such as object unity. Indeed, a somewhat modified network enabled modelling object unity while still accounting for occlusion and object permanence in chapter 3. Furthermore, there was nothing in principle that would have prevented the model to be extended to capture studies on causality as well. For example, additional units representing relative depth of objects would allow the model to build representations of launching in their hidden layers.

At that moment, technical reasons prevented us from doing so: the object unity network has become so large that learning became increasingly difficult. It was already very difficult to account for so many different studies which required a lot of twisting and tweaking and smart choices of the network parameters. In spite of these technical difficulties that reflect the plague of the entire field, it would be short-sighted to assume that otherwise, we were on the right path. We believe that we *were* indeed on the right path but something important was wrong. Specifically, the difficulties of extending the network to capture more phenomena was only a symptom of the ineffectiveness of the representations. In chapter 4 we already discussed that a featurewise representation would make the network more efficient and therefore easier to extend. This hits the heart of the problem: extend-

ing the network by e.g. causality representations would mix the existing hidden layer representations with the new ones. This certainly would destroy the existing representations to some extent because they all would have to be processed by the same hidden layers. This is a general problem in neural network research known as the stability-plasticity dilemma (Grossberg, 1988). Extending the number of neurons obviously only postpones the central problem: even if that works it would require an over-proportional amount of tweaking and twisting in order to get the network working. Intuitively, this kind of work will not pay off because there is not much that we could expect to learn from that venture. In other words, the network does not acquire independent representations that would constitute an efficient strategy for further, open-ended learning. On the other hand, it seems that infants indeed possess such representations as we concluded in chapter 4.

For these reasons it was necessary to look deeper into the question of where infants' understanding of regularities and structures originates from and what guides learning in infancy. Therefore, in the third project we investigated sequence learning and the formation of visual expectations. In a nutshell, the spike timing dependent plasticity mechanism imprinted the structure of the sequence into the network reservoir. The crucial question is whether this is really the process that happens in the infant's brain. On the one hand, this mechanism is consistent with the finding that infants use context-based representations of sequences (Lewkowicz and Berent, 2009), i.e. infants seem to track statistical relations among specific sequence elements (e.g. LR, RL) rather than using ordinal information (e.g. R is third). On the other hand, it is necessary to study the microdevelopment of infants' behavior and learning during the presentation of the sequence. For example, Wentworth and Haith (1998) found that 2- and 3-month-olds require at least 10-20 alternations of the LR sequence in order to shift from repetitive saccades, i.e. eye movements in the same direction as the previous one, in alternating saccades. This may indicate a predisposition of infants to expect objects moving in a continuous and linear fashion (on a small scale). In any case, this branch of research is barely touched and creates an exciting opportunity to investigate how infants learn about regularities in the environment and forecast future events. It will be interesting to test the predictions of the model in collaboration with experimental groups.

It is remarkable that we started with the objective to find the first steps that babies make on their way to understand the world and head on from there to later stages of development. Contrary to that intention we found ourselves going *back* in the age of infants asking rather how infants arrived at the modelled abilities in the first place. For example, modeling causality lead us to the conclusion that perceiving one object launching another cannot be understood without being able to segregate objects in the first place. After modeling this object unity we found ourselves being confronted with even more fundamental problems of how infants generally extract regularities from the environment. Superficially, it looks like we are stepping back in our progress but at a second glance it becomes clear that this is the way scientific investigation takes place and it is beautiful because it takes our understanding deeper and deeper toward the Big Bang of cognition.

References

Azad, P., Gockel, T., and Dillmann, R. (2008). *Computer Vision: Principles and Practice*. Elektor Electronics.

Baillargeon, R. (1986). Representing the existence and the location of hidden objects: Object permanence in 6- and 8-month-old infants. *Cognition*, 23:21–41.

Baillargeon, R. (1987). Object permanence in 3 1/2- and 4 1/2-month-old infants. *Developmental Psychology*, 23(5):655–664.

Baillargeon, R. (1994). How do infants learn about the physical world? *Current Directions in Psychological Science*, 3(5):133–140.

Baillargeon, R. (1999). Young infants' expectations about hidden objects: a reply to three challenges. *Developmental Science*, 2:115–163.

Baillargeon, R., Spelke, E., and Wasserman, S. (1985). Object permanence in five-month-old infants. *Cognition*, 20:191–208.

Balkenius, C. and Johansson, B. (2007). Anticipatory models in gaze control: a developmental model. *Cognitive Processing*, 8(3):167–174.

Benson, J. B., Cherny, S. S., Haith, M. M., and Fulker, D. W. (1993). Rapid assessment of infant predictors of adult IQ: The midtwin-midparent approach. *Developmental Psychology*, 29(3):434–447.

Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming*. Athena Scientific.

Bi, G. Q. and Poo, M. M. (1998). Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of Neuroscience*, 18:10464–10472.

Canfield, R. L. and Haith, M. M. (1991). Infants' visual expectations for symmetric and asymmetric stimulus sequences. *Developmental Psychology*, 27(2):198–208.

Canfield, R. L., Smith, E. G., Brezsnyak, M. P., Snow, K. L., Aslin, R. N., Haith, M. M., Wass, T. S., and Adler, S. A. (1997). Information processing through the first year of life: A longitudinal study using the visual expectation paradigm. *Monographs of the Society for Research in Child Development*, 62(2):1–160.

Canfield, R. L. and Smith, G. E. (1996). Number-based expectations and sequential enumeration by 5-month-old infants. *Developmental Psychology*, 32(2):269–279.

Canfield, R. L., Wilken, J., Schmerl, L., and Smith, E. G. (1995). Age-related change and stability of individual differences in infant saccade reaction time. *Infant Behavior and Development*, 18(3):351 – 358.

Clark, E. V. (2004). How language acquisition builds on cognitive development. *Trends in Cognitive Sciences*, 8(10):472 – 478.

DiLalla, L. F., Thompson, L. A., Plomin, R., Phillips, K., Fagan III, J. F., and Haith, M. M. (1990). Infant predictors of preschool and adult IQ: A study of infant twins and their parents. *Developmental Psychology*, 26(5):759–769.

Doya, K. (1992). Bifurcations in the learning of recurrent neural networks. In *IEEE International symposium on circuits and systems*, volume 6, pages 2777–2780.

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

Frey, P. W. and Ross, L. E. (1968). Classical conditioning of the rabbit eyelid response as a function of interstimulus interval. *Journal of Comparative and Physiological Psychology*, 65(2):246–250.

Funahashi, K. and Nakamura, Y. (1993). Approximation of dynamical systems by continuous time recurrent neural networks. *Neural Networks*, 6(6):801 – 806.

Gilmore, R. O. and Thomas, H. (2002). Examining individual differences in infants' habituation patterns using objective quantitative techniques. *Infant Behavior and Development*, 25:399–412(14).

Gopnik, A., Meltzoff, A. N., and Kuhl, P. K. (1999). *The Scientist in the Crib: Minds, Brains, and How Children Learn.* William Morrow and Company.

Grossberg, S. (1988). Competitive learning: From interactive activation to adaptive resonance. pages 213–250.

Haith, M. M., Hazan, C., and Goodman, G. S. (1988). Expectation and anticipation of dynamic visual events by 3.5-month-old babies. *Child Development*, 59(2):467–479.

Haith, M. M. and McCarty, M. E. (1990). Stability of visual expectations at 3.0 months of age. *Developmental Psychology*, 26(1):68–74.

Hertz, J., Krogh, A., and Palmer, R. G. (1991). *Introduction to the theory of neural computation. Lecture notes volume I.* Perseus Publishing.

Hespos, S. J. and Baillargeon, R. (2001). Reasoning about containment events in very young infants. *Cognition*, 78:207–245.

Hume, D. (1740). A treatise of human nature. Oxford: Clarendon Press.

Jacobson, S. W., Jacobson, J. L., O'Neill, J. M., Padgett, R. J., Frankowski, J. J., and Bihun, J. T. (1992). Visual expectation and dimensions of infant information processing. *Child Development*, 63(3):711–724.

Jaeger, H. (2001). The "echo state" approach to analysing and training recurrent neural networks. *Technical Report GMD Report 148, German National Research Center for Information Technology.*

James, W. (1890). *The Principles of Psychology.* Cambridge, MA: Harvard University Press, 1981.

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

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

Johnson, S. P., Bremner, J. G., Slater, A., Mason, U., Foster, K., and Cheshire, A. (2003). Infants' perception of object trajectories. *Child Development*, 74(1):94–108.

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

Kellman, P. J. and Arterberry, M. E. (1998). *The cradle of knowledge: development of perception in infancy*. MIT Press.

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

Lazar, A., Pipa, G., and Triesch, J. (2007). Fading memory and time series prediction in recurrent networks with different forms of plasticity. *Neural Networks*, 20:312–322.

Lazar, A., Pipa, G., and Triesch, J. (2009). Sorn: a self-organizing recurrent neural network. *Frontiers in Computational Neuroscience*, 3.

Leslie, A. M. (1982). The perception of causality in infants. *Perception*, 11:173–186.

Leslie, A. M. (1994). Tomm, toby, and agency: Core architecture and domain specificity. In Hirschfeld, L. A. and Gelman, S. A., editors, *Mapping the mind: Domain specificity in cognition and culture*. New York: Cambridge University Press.

Lewkowicz, D. J. and Berent, I. (2009). Sequence learning in 4-month-old infants: Do infants represent ordinal information? *Child Development*, 80(6):1811–1823.

Lin, L. J. and Mitchell, T. M. (1992). *Memory approaches to reinforcement learning in non-markovian domains*. Technical Report CMU-CS-92-138, Carnegie Mellon University, Pittsburgh, PA.

Lukosevicius, M. and Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149.

Maass, W., Natschlger, T., and Markram, H. (2002). Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Computation*, 14(11):2531–2560.

Mandler, J. M. (1992). How to build a baby: II. Conceptual primitives. *Psychological Review*, 99(4):587–604.

Mandler, J. M. (2004). *The foundations of mind. Origins of conceptual thought*. Oxford Series in Cognitive Development.

Mandler, J. M. and McDonough, L. (1993). Concept formation in infancy. *Cognitive Development*, 8(3):291–318.

Mandler, J. M. and McDonough, L. (1998). On developing a knowledge base in infancy. *Developmental Psychology*, 34(6):1274–1288.

Marcus, G. F. (2001). Plasticity and nativism: Towards a resolution of an apparent paradox. In Wermter, S., Austin, J., and Willshaw, D., editors, *Emergent neural computational architectures based on neuroscience*, pages 368–382. Heidelberg: Springer.

Mareschal, D., French, R. M., and Quinn, P. C. (2000). A connectionist account of asymmetric category learning in early infancy. *Developmental psychology*, 36(5):635–645.

Mareschal, D. and Johnson, S. P. (2002). Learning to perceive object unity: a connectionist account. *Developmental Science*, 5:151–185.

Mareschal, D., Plunkett, K., and Harris, P. (1999). A computational and neuropsychological account for object-oriented behaviors in infancy. *Developmental Science*, 2:306–317.

Markram, H., Lübke, J., Frotscher, M., and Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science*, 275(5297):213–215.

McClelland, J. L. (1995). A connectionist perspective on knowledge and development. In Simon, T. J. and Halford, G. S., editors, *Developing cognitive competence: New approached to process modeling*. Hillsdale, NJ: Lawrence Erlbaum.

Mervis, C. B. (1987). Child-basic object categories and early lexical development. In Neisser, U., editor, *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization*. Cambridge University Press.

Michotte, A. (1963). *The perception of causality*. New York: Basic Books.

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

Munakata, Y. and Stedron, J. M. (2002). Modeling infants' perception of object unity: what have we learned? *Developmental Science*, 5:173–180.

Orban, G. A., Kennedy, H., and Bullier, J. (1986). Velocity sensitivity and direction selectivity of neurons in areas V1 and V2 of the monkey: influence of eccentricity. *Journal of Neurophysiology*, 56:462–480.

Pauen, S. (2002). The global-to-basic level shift in infants' categorical thinking: First evidence from a longitudinal study. *International Journal of Behavioral Development*, 26(6):492–499.

Piaget, J. (1971). *Biology and knowledge*. Chicago: University of Chicago Press.

Rao, R. P. N. and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2:79–87.

Reznick, J. S., Chawarska, K., and Betts, S. (2000). The development of visual expectations in the first year. *Child Development*, 71(5):1191–1204.

Rolls, E. and Deco, G. (2002). *Computational Neuroscience of Vision*. Oxford University Press, USA.

Rosch, E. and Mervis, C. B. (1975). Familiy resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7:573–605.

Rosch, E., Mervis, C. B., Gray, W., Johnson, D., and Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 3:382–439.

Rose, S. A., Feldman, J. F., Jankowski, J. J., and Caro, D. M. (2002). A longitudinal study of visual expectation and reaction time in the first year of life. *Child Development*, 73(1):47–61.

Rovee-Collier, C. (1997). Dissociations in infant memory: Rethinking the development of implicit and explicit memory. *Psychological Review*, 104(3):467–498.

Rovee-Collier, C. (1999). The development of infant memory. *Current Directions in Psychological Science*, 8(3):80–85.

Saeb, S., Weber, C., and Triesch, J. (2009). Goal-directed learning of features and forward models. *Neural Networks*, 22(5-6):586 – 592. Advances in Neural Networks Research: IJCNN2009, 2009 International Joint Conference on Neural Networks.

Saffran, J. R., Aslin, R. N., and Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

Schlesinger, M. and Young, M. E. (2003). Examining the role of prediction in infants' physical knowledge. In *Proceedings of the Twenty-fifth Annual Conference of the Cognitive Science Society (Boston, MA, USA)*, pages 1047–1052.

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

Slater, A., Morison, V., Town, C., and Rose, D. (1985). Movement perception and identity constancy in the new-born baby. *British journal of developmental psychology*, 3:211–220.

Sokolov, E. N. (1963). *Perception and the conditioned reflex*. Oxford: Pergamon.

Sokolov, E. N. (1975). Neuronal mechanisms of the orienting reflex. In Sokolov, E. N. and Vinogradova, O. S., editors, *Neuronal mechanisms of the orienting reflex*. Hillsdale, NJ: Lawrence Erlbaum.

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

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

Spelke, E. S. and Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, 10(1):89–96.

Spencer, J. P., Blumberg, M. S., McMurray, B., Robinson, S. R., Samuelson, L. K., and Tomblin, J. B. (2009). Short arms and talking eggs: Why we should no longer abide the nativist-empiricist debate. *Child Development Perspectives*, 3(2):79–87.

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

Szita, I., Gyenes, V., and Lrincz, A. (2006). Reinforcement learning with echo state networks. In *Artificial Neural Networks ICANN 2006*, pages 830–839. Springer Berlin / Heidelberg.

Tesauro, G. and Sejnowski, J. J. (1989). A parallel network that learns to play backgammon. *Artificial Intelligence*, 39(3):357–390.

Vinogradova, O. S. (1975). The hippocampus and the orienting reflex. In Sokolov, E. N. and Vinogradova, O. S., editors, *Neuronal mechanisms of the orienting reflex*. Hillsdale, NJ: Lawrence Erlbaum.

von Hofsten, C., Kochukhova, O., and Rosander, K. (2007). Predictive tracking over occlusions by 4-month-old infants. *Developmental Science*, 10:625–640.

Wentworth, N. and Haith, M. M. (1998). Infants' acquisition of spatiotemporal expectations. *Developmental Psychology*, 34(2):247–257.

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt. *Psychologische Forschung*, 4:301–350.

Zigmond, M. J., Bloom, F. E., Landis, S. C., Roberts, J. L., and Squire, L. R. (2003). *Fundamental Neuroscience*. Elsevier Science, USA.

Zusammenfassung der Arbeit

Dieses Kapitel beinhaltet eine deutsche Zusammenfassung der Arbeit. Bei Fachbegriffen, zu denen es keine deutschen Entsprechungen gibt, wurden die englischen Bezeichnungen beibehalten.

Die vorliegende Dissertation untersucht die Entwicklung früher kognitiver Fähigkeiten im Säuglingsalter mit neuronalen Netzen. Grundlegende Ereignisse in der visuellen Wahrnehmung wie durch Stöße verursachte Bewegung, Verdeckung, Objektpermanenz, Verfolgen bewegter Objekte hinter Verdeckungen, Wahrnehmung von Objekteinheit und das Erlernen von Reizfolgen werden in einem vereinheitlichenden, theoretischen Rahmen modelliert, während die Nähe zu experimentellen Ergebnissen der Entwicklungspsychologie im Säuglingsalter gewahrt wird.

Entwicklung der Wahrnehmung von Kausalität und Verdeckung

Die Debatte um Kausalität kann bis Hume (1740) zurückverfolgt werden, in dessen klassischer Abhandlung die Wahrnehmung von Kausalität in einfachen mechanischen Ereignissen das Resultat von wiederholter Wahrnehmung konstanter Verknüpftheit zweier Ereignisse ist. Michotte (1963) argumentierte, dass Kausalität direkt
wahrgenommen werden könne, während Leslie (1994) davon ausging, dass ein angeborenes Kraft- oder Druckkonzept vonnöten sei. Mandler (2004) schlug vor,

dass die Betrachtung des Übertrags von Bewegung für Kinder eine ausreichende Grundlage für eine frühe Interpretation kausaler physikalischer Ereignisse sei.

Ein anderes grundlegendes Verständnis aus dem Bereich naiver Physik stellt die Wahrnehmung von Verdeckung dar. Die Tatsache, dass Objekte, auch wenn sie hinter einer Verdeckung verschwinden, weiterhin existieren — die sog. Objektpermanenz — ist bei Säuglingen nicht von Geburt an vorhanden, sondern entwickelt sich um das Alter von vier Monaten herum (Baillargeon et al., 1985, Baillargeon, 1987). Auch das Nachverfolgen bewegter Objekte nach ihrer Verdeckung, d.h. ihre fortgesetzte Repräsentation im Gehirn, ist eine Fähigkeit, die sich auch erst nach einigen Monaten nach der Geburt entwickeln muss (Johnson et al., 2003, von Hofsten et al., 2007).

Die meisten bisherigen Experimente und Modelle untersuchen diese Fähigkeiten nur getrennt voneinander, während ein vereinheitlichtes Modell noch ausblieb. Wir präsentieren ein künstliches neuronales Netz, das zwei Experimente zur Wahrnehmung von Kausalität und Verdeckung in einem einheitlichen Rahmen modelliert und erklärt. Es handelt sich um ein einfaches rekurrentes Netz, das sog. Elman-Netz (Elman, 1990), das dazu trainiert wird, Reizabfolgen, die seiner Eingabeschicht präsentiert werden, vorherzusagen. Bei der Eingabe handelt es sich um bewegte Pixel, die sich entweder linear bewegen, sich gegenseitig verdecken oder stoßen. Ein wichtiges, von uns in der Literatur zum ersten mal eingeführtes Element, ist das Vortrainieren des Netzes, das je nach Länge des Trainings das Alter des Säuglings modelliert. Die Eingabereize bilden demnach die visuelle Erfahrung des Säuglings im Laufe des Lebens ab. Nach dem Vortraining wird das Netz analog zu den tatsächlich durchgeführten Experimenten wiederholt Verdeckungs- und Stoßabfolgen ausgesetzt. Der Backpropagation-Lernalgorithmus führt dazu, dass der Vorhersagefehler dieser Abfolgen kontinuierlich reduziert wird, analog zu den gemessenen mittleren Betrachtungszeiten der Säuglinge auf die ihnen gezeigten Reize, wie es im sog. Habituationsparadigma beobachtet wird. Dieser Umstand erlaubt die Modellierung von Blickzeiten mithilfe des Vorhersagefehlers des Netzes, zudem auch bei Säuglingen davon ausgegangen wird, dass Vorhersagefehler künftiger Ereignisse für eine erhöhte Aufmerksamkeit und damit für ausgiebigere Betrachtung der Reize verantwortlich sind (Vinogradova, 1975, Sokolov, 1963,

1975, Gilmore and Thomas, 2002).

Das Netz erklärt die Wahrnehmung von Kausalität und Verdeckung auf eine vereinheitlichte Weise, in dem es beide Phänomene auf den Umstand zurückführt, dass sie durch das Erlernen von statistischen Regelmäßigkeiten der Reizabfolgen repräsentiert werden können. Insbesondere konnte das Blickverhalten der Säuglinge sowohl bei den verschiedenen Reizen als auch im Verlauf ihrer Entwicklung dadurch detailliert mithilfe desselben Netzes erklärt werden.

Entwicklung der Wahrnehmung von Objekteinheit, Objektpermanenz und Verdeckung

Im zweiten Modell wird der erarbeitete theoretische Rahmen zu einem größeren auf Vorhersage trainierten Netz erweitert, das die Entwicklung der Wahrnehmung von Objekteinheit, Objektpermanenz und Verdeckung im Säuglingsalter modelliert. Das Verständnis, dass die Welt aus Objekten besteht, deren Bestandteile sich kohärent und auf stetigen Pfaden bewegen und die fortwährend existieren, auch wenn sich verdeckt werden, ist ein wichtiger Schritt in unserer ontogenetischen Entwicklung. Zum Beispiel hat die Erforschung der Wahrnehmung von Objekteinheit gezeigt, dass vier Monate alte Säuglinge in der Lage sind, einen sich hinter einem Quader bewegenden Stab als vollständig wahrzunehmen, obwohl lediglich die beiden herausschauenden Stabenden sichtbar sind (Kellman and Spelke, 1983, Johnson and Aslin, 1995, Johnson and Náñez, 1995).

Nebst Objekteinheit werden die Phänomene der Verdeckung und Objektpermanenz aus dem vorherigen Kapitel aufgegriffen. Es wird gezeigt, dass diese verschiedenen modellierten Phänomene in einem einzigen theoretischen Rahmen vereinheitlicht werden können und damit experimentelle Ergebnisse aus 14 Säuglingsstudien erklärt werden können. Die Modelle demonstrieren, dass diese Entwicklungsphänomene erklärt werden können, indem statistische Regelmäßigkeiten in der visuellen Umgebung repräsentiert und vorhergesagt werden. Die Modelle

nehmen an, dass (1) verschiedene neuronale Populationen im visuellen Kortex des neugeborenen Säuglings auch verschiedene Bewegungsrichtungen der visuellen Reize verarbeiten, was durch neurowissenschaftliche Belege untermauert ist (Orban et al., 1986), und (2) Lernmechanismen vorhanden sind, die zur Vorhersage künftiger Ereignisse führen. Insbesondere demonstrieren die Modelle, dass keine angeborenen Kraftkonzepte, Module zur Bewegungsanalyse (Leslie, 1994), Detektoren korrelierter Bewegung (Mareschal and Johnson, 2002), besondere Wahrnehmungsregeln oder die Fähigkeit über Entitäten "nachzudenken" (Spelke, 1994), wie sie weitverbreitet in der Literatur der Entwicklungspsychologie zu finden sind, notwendig sind, um die diskutierten Phänomene zu erklären.

Entwicklung visueller Erwartungen und das Erlernen von Reizabfolgen

Da sich die Vorhersage künftiger Ereignisse für die theoretische Erklärung diverser Entwicklungsphänomene und als Leitfaden für das Lernen im Säuglingsalter als fruchtbar erwiesen hat, spricht das dritte Modell die Entwicklung visueller Erwartungen selbst an. Eine Reihe von Studien haben gezeigt, dass selbst wenige Monate alte Säuglinge in der Lage sind, Abfolgen von links und rechts auf dem Bildschirm auftauchenden Bildern durch antizipatorische Augenbewegungen vorherzusagen (Haith et al., 1988, Canfield et al., 1997, Wentworth and Haith, 1998, Jacobson et al., 1992, Canfield and Haith, 1991, Canfield and Smith, 1996, Rose et al., 2002, Reznick et al., 2000). Außerdem wurde in den genannten Studien u.a. ein Abfall der Reaktionszeiten im Laufe der Experimente und mit steigendem Alter der Säuglinge beobachtet.

Ein selbstorganisierendes, vollständig rekurrentes, neuronales Netz, das interne Repräsentationen von Eingabeabfolgen formt und sie auf Augenbewegungen abbildet, wird vorgeschlagen, um den obigen Datensatz zu modellieren. Die für das Verstärkungslernen verantwortliche Architektur des Modells lernt, antizipatorische Augenbewegungen verschiedener Reizabfolgen auszuführen. Das Modell postuliert,

dass das Blickverhalten von Säuglingen vom Ziel, die Betrachtungszeit interessanter Reize zu maximieren, geleitet ist, und erklärt damit das Auftauchen und die Entwicklung antizipatorischer Augenbewegungen und Reaktionszeiten. Im Gegensatz zu herkömmlichen Modellen mit neuronalen Netzen in der Entwicklungspsychologie benutzt das Modell lokale Lernregeln und enthält mehrere biologisch plausible Elemente wie exzitatorische und inhibitorische Neuronen, spike-timing dependent plasticity (STDP), intrinsische Plastizität (IP) und synaptische Skalierung. Es bedient sich eines dynamischen, rekurrenten Reservoirs, das sich nachweislich für Vorhersageaufgaben gut eignet und herkömmliche Netze mit statischem Reservoir in ihrer Leistung übersteigt (Lazar et al., 2009). Wir erweitern dieses Reservoir mit einer Architektur für das Verstärkunglernen, womit das Modell auch aus technischer Sicht neuartig ist.

Das gesamte Netz erlernt die Reizabfolgen, z.B. eine links-links-rechts Folge, indem es diese Abfolge der Eingabeneuronaktivitäten nichtlinear auf die Aktivität des gesamten Reservoirs abbildet und damit in einen hochdimensionalen Reservoirzusandsraum projiziert. Die damit erreichte sog. lineare Trennbarkeit der Reservoirzustände bildet die Basis für das Auslesen der Zustände mithilfe der Architektur für das Verstärkungslernen. Dieses aus dem sog. reservoir computing abgeleitete Prinzip wird nun durch neuronale Plastizität des Reservoirs erweitert, das noch bessere Repräsentationen der Eingabefolgen erreicht und eine Grundlage für die Modellierung verschiedener Altersstufen bildet. Eine Analyse der Wechselwirkung zwischen dem Reservoir und dem Verstärkungslernen ergibt einen nichttrivialen, dynamischen Zusammenhang, der in der Analyse untersucht wird.

Das Modell erklärt zwölf experimentelle Studien und sagt unter anderem das Antizipationsverhalten von Säuglingen für beliebige Reizabfolgen und den erleichterten Wiedererwerb bereits erlernter Abfolgen vorher. Durch die Fähigkeit selbst identische, wiederholte Eingaben auf verschiedene Reservoirzustände abzubilden, wird es dem Modell möglich, identische Eingabefolgen abzuzählen, was als Grundlage für die Erklärung der Entwicklung numerischer und mathematischer Fähigkeiten bei Säuglingen dienen kann.

Alle Modelle betonen die Entwicklung der Wahrnehmung der diskutierten Phänomene und erklären, wie und warum Veränderung während der Entwicklung stattfindet — Fragen, die auf experimentelle Weise schwierig zu klären sind. Trotz der Verschiedenheit der diskutierten Phänomene beruhen alle drei Projekte auf dem selben Prinzip: die Vorhersage künftiger Ereignisse. Dieses Prinzip postuliert, dass die kognitive Entwicklung im Säuglingsalter zu großen Teilen vom Aufbau interner Modelle und Repräsentationen der visuellen Umgebung und vom Benutzen dieser Modelle zur Vorhersage der künftigen Entwicklung dieser Umgebung geleitet ist.

Curriculum Vitae

Arthur Franz

Frankfurt Institute for Advanced Studies (FIAS)
Ruth-Moufang-Str. 1
60438 Frankfurt am Main
Phone: +49 69 79847612
Email: franz@fias.uni-frankfurt.de

Personal

Date of birth	January 26, 1982
Place of birth	Ufa, Russia
Nationality	German

Education

2007 - 2010	PhD student at the Frankfurt Institute of Advanced Studies (FIAS). Topic: "Computational models of cognitive development in infancy". Research group: Prof. Dr. Jochen Triesch.
2003 - 2006	Diploma in physics at the universities of Erlangen-Nürnberg and Regensburg. Topic of diploma thesis: "Evolution of regulated chemical networks". Average grade: 1.18
2002 - 2003	Military service in Roth, Germany.
1992 - 2002	Gymnasium in Weißenburg and Dinkelsbühl. Average grade: 1.4
1989 - 1992	Middle school in Koktschetav, Kasachstan.

Scholarships

2004 - 2006	Scholarship from the Studienstiftung des deutschen Volkes Ranked among the top 1% of German students.
2004 - 2006	Member of the Elite Network of Bavaria. Participation in the Elite Graduate Programme in Physics.

International experience

08/2008	International Conference on Development and Learning, Monterey, USA.
10/2007	Coherent Behavior in Neural Networks, Mallorca, Spain.
06/2007	International Conference on Development and Learning, Imperial College London, UK.
02/2007	Conference on cortical plasticity in Amsterdam.
09 - 12/2006	Research in the laboratory of Prof. Nancy Kanwisher at the Massachussetts Institute of Technology (MIT), Cambridge, USA. fMRI study on color frequency sensitivity of the visual cortex and psychophysical experiments on visual attention.

Workshops and summer schools

08/2007	Summer School in Theoretical Neuroscience at FIAS
05/2007	Meeting of the Philosophy of Mind. Topic: "Neuroethics"
08/2006	Summer school on evolutionary psychology, organized by the Studienstiftung des deutschen Volkes

Publications

A. Franz, T. Kolling and J. Triesch (2010). A computational model of the development of visual expectation in infancy. *Submitted to Developmental Science.*

A. Franz and J. Triesch (2010). A computational model of the development of visual expectation in infancy. *Poster presented at the International Conference on Infant Studies.*

A. Franz and J. Triesch (2009). A unified computational model of the development of object unity, object permanence, and occlusion perception. *Infant Behavior and Development. In review.*

A. Franz and J. Triesch (2008). Modeling the development of causality and occlusion perception in infants. *Proceedings of the 7th International Conference on Development and Learning*, 174-179.

A. Franz and J. Triesch (2007). Emergence of Disparity Tuning during the Development of Vergence Eye Movements. *Proceedings of the 6th International Conference on Development and Learning*, 31-36.

Other skills and competences

Languages	Russian (native), German (fluent), English (fluent), French (good)
Presentation skills	Former member of "Toastmasters International"
Programming skills	Java, Matlab, HTML, Java script, Prolog