MINDREADERS IN THE CRIB:

COGNITIVE MECHANISMS FOR REPRESENTING OTHERS' MENTAL STATES IN HUMAN INFANTS

DORA KAMPIS

SUPERVISOR: ÁGNES M. KOVÁCS
SECONDARY SUPERVISOR: GERGELY CSIBRA

Thesis submitted for the degree of Doctor of Philosophy (PhD)

Department of Cognitive Science
Central European University

Submitted: Budapest, September 2016
Defended: January 9, 2017

Originality Statement

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at Central European University or any other educational institution, except where due acknowledgment is made in the form of bibliographical reference.

I also declare that the intellectual content of this thesis is the product of my own work, and the claims here reflect my own thinking. However, as I believe scientific research is inherently a social activity, and all work receives support from others in theoretical, methodological, or stylistic matters; I will present this work in first-person plural voice.

The present thesis includes work that appears in the following papers:

Kampis, D., Parise, E., Csibra, G., & Kovács, Á. M. (2015). Neural signatures for sustaining object representations attributed to others in preverbal human infants. *Proceedings of the Royal Society B: Biological Sciences*, 282(1819).

Kampis, D., Parise, E., Csibra, G., & Kovács, Á. M. (2016). On potential ocular artefacts in infant electroencephalogram: a reply to comments by Köster. *Proceedings of the Royal Society B: Biological Sciences*, 283(1835).

Dora Kampis

Abstract

A crucial part of human cognition is to understand that people are guided not just by external factors, but also by their mental states. This capacity, termed "Theory of Mind", has been of great interest in the past four decades to researchers form a variety of fields. A pressing question is how the ability to form metarepresentations of others' mental states develops, and whether it is present in human infants. The present thesis investigated the cognitive mechanisms that may enable young infants to represent other people's mental representations. The first two experiments explored the neuro-cognitive bases of infants' ability to encode the world from another person's perspective, hypothesizing that the cognitive systems involved in representing the world from infants' own perspective are also recruited for encoding others' beliefs. Indeed, there was a common neural activation when infants sustained an object representation, and when they could attribute such representations to someone else. Three subsequent studies investigated infants' abilities to ascribe to others beliefs based on correct or mistaken individuation of objects using spatiotemporal or feature/kind information; and found that infants can represent others' beliefs involving multiple objects, and object identity. Finally, the last two experiments probed the flexibility of infants' mental state attributions by testing how infants can integrate new information into their already existing representations. Together, these studies point to the possibility of an early developing, flexible, and powerful apparatus suitable to handle multiple concurrent representations; which may be the core of a mature mindreading ability in adulthood.

Acknowledgements

I would like to thank my supervisor, Ágnes M. Kovács, for taking me as her student and helping me through all these years. Ági, I could not have wished for better mentoring and support, it is a privilage to have worked with you in the past 5 years. I greatly admire your way of thinking and I hope I have adopted some of it through this time. Working with me may be challenging at times; thank you for your patience and persistence. It will be difficult to part, and I hope we will continue to work together in the future.

I am grateful for the guidance of Gergely Csibra, my secondary supervisor who among all his duties always had time to answer my questions. Gergő, thank you for the calm rationality you bring into my somewhat hectic mind. You set a great example for scientific conduct and how to lead a lab, which I hope to have learnt from.

Furthermore, I want to thank the entire faculty and staff of the Department of Cognitive Science. We started as the first generation of graduate students, and it was a learning process for all. But I felt welcome and with great support and flexibility from the day I first set foot at CEU. Thanks to all the faculty who make this department such a great intellectual hub. Thanks to Réka, Andrea, and Ági who help our lives in so many ways it could not be listed here. And without the help of the Babylab staff our research would simply not be possible. Thank you all for being so great, flexible, precise and fun to work with, even at difficult times. You are amazing.

I owe special thanks to every member of the Cognitive Development Center. Rubeena, Ágota, Denis; I am glad we were there for each other in the crazy times of the first year. Ruby; I hope one day I will get to visit the Babylab you set up in India (or elsewhere). I appreciate our friendship and hope to keep it wherever we will end

up. Mikołaj, thank you for our pact of sarcasm, and for believing in me even when I didn't. I could always turn to you for support, and you had a great role in my personal and professional growth. Ernő, you taught me not to always reply to what others say; and you are one of the main reasons it is fun to work here. Olivier, you brought good taste and humble elegant argumentation to the lab. Eugenio, thanks for making EEG look less intimidating. Gyuri, thank you for the inspiring conversations. Carina, I am glad we shared a few months of our PhD lives. Frances, you were the best office mate! Katka, I am glad you came into our lives with your sharp mind and tongue, and am extremely thankful for all the great advice and support you have given me throughout the writing process. And thanks to everyone else who makes working here such a pleasure.

I have benefited greatly from many conversations over the years, at the department and at various scientific meetings. I am sure I may leave out some names and to those I apologize. I remember chats with Dan Sperber, Alan Leslie, Steve Butterfill, Pierre Jacob, Dana Samson, Ian Apperly, Hannes Rakoczy, Vicky Southgate, Susan Carey, Liz Spelke, Lisa Feigenson, Justin Halberda, Mohinish Shukla, Paula Rubio-Fernandez. The fact that I applied to this program was thanks to Ildikó Király, who was my Master's supervisor and my point of entry into science. Ildikó, you helped me take off on this path, and I am thankful for the encouragement.

Throughout my years as a graduate student, I was supported by various CEU funding sources: By CEU's Doctoral Scholarship, Write-Up Grant, and an academic achievement award I had received. My visit at Johns Hopkins was supported by CEU's Doctoral Research Support Grant.

I am thankful for the help of my family and all friends, who supported me through the many years that led here. My father, who is so critical it pushes me to aim for better. My mother, who drives across Europe when needed to give us support. My sister, who is always there as a secret ally. Having you as background enabled me to pursue my interests. Anna, Saci, Zsofi: thanks for being my friends through good and bad.

Finally, I am thankful for all the support and love from my husband, Bence. Thank you for learning to understand my passion and bringing stability and so much joy into my life. I am excited about our future adventures together.

I dedicate this thesis to my grandparents. I am sure you would be proud.

TABLE OF CONTENTS

ORIGINALITY STATEMENT **3**

ABSTRACT **5**

ACKNOWLEDGEMENTS **7**

INTRODUCTION **17**

Research aims 20

CHAPTER 1 ..... **23**

1.1. THEORY OF MIND: DEVELOPMENTAL FINDINGS AND THEORIES **28**

1.1.1. Assessing ToM development: empirical findings 30

Findings from elicited tasks *30*

Findings from spontaneous measures *34*

1.1.2. Explaining ToM development: theoretical approaches 36

Gradual development accounts *37*

Conceptual change accounts *42*

Mental file theory 45

A two-system approach: the minimal-ToM proposal 51

Critiques of the minimalist account 55

Summary: possible limitations of early ToM abilities *60*

1.2. MECHANISMS OF EARLY TOM ABILITIES **62**

Proposing a one-system ToM with multiple components *62*

1.2.1. Meta-representing belief contents in infants 69

Indicators of shared mechanisms for own and attributed representations *71*

The modulation effect: How belief contents influence one's own behavior *75*

Object representation capacities in service of belief representations *79*

Summary: cognitive underpinnings of representing belief contents in infants *87*

1.2.2. Automaticity and flexibility of ToM mechanisms 88

Belief ascription in perspective taking *89*

Updating and discarding mental states *92*

1.3. SUMMARY AND OUTLINE OF PRESENT WORK **96**

CHAPTER 2 101

NEURAL UNDERPINNINGS OF REPRESENTING OTHERS' BELIEFS IN INFANTS..... 101

Representing attributed object representations..... 103

2.1. EXPERIMENT 1 108

2.1.1 Materials and Methods..... 109

Participants..... 109

Procedure..... 109

Stimuli..... 109

EEG recording and analysis..... 111

2.1.2 Results..... 114

2.1.3 Discussion..... 116

2.2. EXPERIMENT 2 117

2.2.1 Materials and Methods..... 118

Participants..... 118

Stimuli..... 118

EEG recording and analysis..... 119

2.2.2. Results..... 120

2.2.3 Additional Analyses (Experiment 1 & 2)..... 122

Late burst activation..... 122

Ruling out potential ocular artifacts..... 124

2.2.4. Discussion..... 129

2.3. GENERAL DISCUSSION 130

CHAPTER 3 133

ATTRIBUTION OF BELIEFS INVOLVING TRACKING MORE THAN ONE OBJECT 133

3.1 INTRODUCTION..... 135

Tracking multiple objects from first-person perspective..... 135

Tracking multiple entities in belief representations..... 137

Continuous measures grasp modulation of one's own behavior by others' beliefs..... 140

3.2. EXPERIMENT 3: REPRESENTING OTHERS' BELIEFS ABOUT MULTIPLE OBJECTS 142
3.2.1 Methods 144
Participants *144*
Materials *144*
Procedure *145*
Familiarization trials 146
Test Trials: FB1 condition 147
Test Trials: FB0 condition 149
Coding *149*
Search duration 149
Pointing and reaching 150
3.2.2. Results 151
Manual search *151*
Pointing and reaching *155*
3.3. DISCUSSION 156

CHAPTER 4 165

ATTRIBUTING BELIEFS INVOLVING INDIVIDUATION OF OBJECTS BASED ON FEATURE/KIND PROPERTIES 165

4.1. INTRODUCTION 167
Individuation based on feature/identity information from first-person perspective 168
Representing beliefs involving individuation based on feature/identity information 170
4.2. EXPERIMENT 4: ATTRIBUTING FALSE BELIEFS BASED ON CORRECT INDIVIDUATION BY FEATURES OF OBJECTS 171
4.2.1. Methods 173
Participants *173*
Materials and Procedure, Coding *173*
4.2.2. Results 175
4.2.3. Discussion 177

4.3. EXPERIMENT 5 – ATTRIBUTING FALSE BELIEFS BASED ON INCORRECT INDIVIDUATION BY FEATURES OF OBJECTS 180
Tracking beliefs about identity of objects 182
4.3.1. Methods 188
Participants 188
Materials 189
Procedure 190
4.3.2. Results 195
4.3.3. Discussion 200
4.4. CONCLUSIONS 206

CHAPTER 5 209

FLEXIBILITY OF INFANTS' AGENCY- AND GOAL ATTRIBUTION 209

5.1 INTRODUCTION 211
5.2. EXPERIMENT 6 216
5.2.1. Methods 218
Participants 218
Procedure 219
Stimuli 219
5.2.2. Results 223
5.2.3. Discussion 224
5.3. EXPERIMENT 7 226
5.3.1. Methods 229
Participants 229
Procedure 229
Stimuli 230
5.3.2. Results 235
5.3.3. Discussion 238
5.4. GENERAL DISCUSSION 241

CHAPTER 6 245

GENERAL DISCUSSION AND CONCLUSIONS..... 245

6.1 SUMMARY OF FINDINGS..... 249

6.2. COMMON FORMAT OF OWN AND ATTRIBUTED REPRESENTATIONS..... 255

Theoretical implications..... 255

Differentiation of own and attributed representations..... 257

6.3. THE MODULATION EFFECT..... 260

Theoretical implications..... 260

Relationship of the modulation effect with other capacities..... 263

6.4. LIMITS OF EARLY TOM 264

Registrations and mental files..... 264

Can infants ascribe mental states involving approximate magnitudes?..... 266

Can infants attribute individuation?..... 267

6.5. FLEXIBILITY OF TOM..... 269

On-line computations of mental states..... 269

What are the rules governing mental state computations?..... 270

6.6. CONCLUSIONS 271

REFERENCES..... 273

Introduction

Humans live in complex social environments, and successful navigation and efficient interactions hinge on our understanding how another person's actions are guided by their unique psychological states. While various nonhuman species show precursors of the ability to encode what their conspecifics see, know or believe (Bugnyar & Heinrich, 2005; Martin & Santos, 2016; Tomasello, Call, & Hare, 2003), humans seem to perform uniquely sophisticated inferences about other people's mental states (Herrmann, Call, Hernàndez-Lloreda, Hare, & Tomasello, 2007). However, it is still debated how these possibly human-specific capacities, termed Theory of Mind (Premack & Woodruff, 1978), develop.

Theory of Mind (ToM) entails the ability to represent others' mental states, such as goals, beliefs, preferences, or desires. Much of research on ToM has been trying to answer the question when the understanding of epistemic states, specifically beliefs, emerges in phylogenetic and ontogenetic development, and what is the representational machinery that enables such cognition. A debated issue is whether human infants are capable of forming metarepresentations of other's mental states in the form of propositional attitudes (Leslie, 2000a; Perner, Leekam, & Wimmer, 1987), and relatedly, whether infant capacities share their core characteristics with adults, or they undergo a radical conceptual change before preschool age (Butterfill & Apperly, 2013; Carey & Johnson, 2000; Carruthers, 2013; Gopnik & Astington, 1988).

A group of proposals argue that organisms can behave in a way that approximate a metarepresentational understanding of other minds, but the underlying cognitive

systems lack some of the core characteristics of fully elaborate ToM. In human infants, such limitations are suggested to be either due to the lack of maturation of more domain-general abilities (Perner, Mauer, & Hildenbrand, 2011), or a separate system operating with simpler and cognitively less demanding representational formats (Apperly & Butterfill, 2009), that is limited in the scope of mental contents it can handle. Similar mechanisms were suggested to operate when people engage in fast, spontaneous, or implicit computations of mental states (Samson, Apperly, Braithwaite, Andrews, & Bodley Scott, 2010).

Others have disputed the necessity of distinct mechanisms for early (or simple) and later (or complex) ToM; arguing that (i) development can be explained through a gradual enrichment of both ToM-related and other cognitive abilities (Carruthers, 2015b; Leslie, 2000a), and (ii) children's difficulties in certain situations, as well as adults' occasional lack of spontaneous mindreading can be explained by external factors such as the additional cognitive demands of the tasks used for assessment of such abilities (Baillargeon, Scott, & He, 2010), pragmatic factors (Helming, Strickland, & Jacob, 2014; Westra, 2016), or participants' covert or overt goals (Carruthers, 2015a; Elekes, Varga, & Király, 2016). These claims are supported by findings suggesting a developmental continuity between early and later ToM abilities (Sodian et al., 2016; Thoermer, Sodian, Vuori, Perst, & Kristen, 2012; Wellman, Phillips, Dunphy-Lelii, & LaLonde, 2004), and a common neural basis of spontaneous and elicited mindreading (Hyde, Aparicio Betancourt, & Simon, 2015; Kovács, Kühn, Gergely, Csibra, & Brass, 2014).

An apparent contradiction seems to be between the speed and flexibility of Theory of Mind cognition. Because of the variety of phenomena it may be involved in,

from rapid interactions, through communication, to jurors' deliberations; calculating mental states is required to potentially happen fast and efficiently, but also to cover a wide range of mental state contents and incorporate extensive information, which may be flexible but effortful. Others have disputed the validity of a speed-flexibility tradeoff in ToM processes, arguing that efficiency may happen through the integrated functioning of multiple separate, specialized, efficient cognitive systems (Carruthers, 2013; Christensen & Michael, 2016; Scholl & Leslie, 1999). One suggestion that aims to offer a solution on how representing others' beliefs may occur in a both fast and efficient manner is the construct of belief file (Kovács, 2015). Belief files are representational constructs that have two placeholders; one for an agent (or mental state-holder) and one for the content. These variables can be separately accessed and manipulated, which enables rapid calculation and modification of the information contained in the file.

Such accounts positing an early developing ToM system face two kinds of challenges. First, there is a need to empirically investigate predictions that follow from how ToM processes might operate. For instance, if particular domains are involved in representing mental states, the signatures of such processes should accompany ToM cognition as well. Second, the predictions from theories suggesting specific limitations of infant mindreading need to be addressed. As the simpler mindreading system cannot grasp the aspectuality of beliefs (i.e. to represent one object under two aspects), representing beliefs about quantifiers or identity should not take place spontaneously, or in infants.

Research aims

The present work investigates the mental representations and cognitive processes underlying Theory of Mind in infants. It explores the possibility that ToM abilities in their core format are present from early on. On one hand, it aims to give an account on the possible mechanisms through which belief contents are handled. On the other hand, it outlines yet unexplored aspects of mindreading with regard to the flexibility of the underlying processes.

This thesis expands on the idea that representing mental states develops early, and entails the orchestrated functioning of many different cognitive systems. We propose that an efficient way to calculate the content variable of belief files is through recruiting one's cognitive apparatus that subserve representing the environment from their own perspective. The studies discussed in Chapter 2 investigated this possibility with electrophysiological methods. We measured neural activation when 8-month-old infants sustain an object representation from their own perspective and when they could attribute such sustained representation to another person. The purpose of these studies was to assess whether the two have similar electrophysiological correlates, which would be suggestive of shared underlying mechanisms. A possible common format of own and attributed representations led to interest in a phenomenon, here referred to as the modulation effect, whereby one's own actions are modulated by others' beliefs (Kovács, Téglás, & Endress, 2010; Samson et al., 2010). The following chapters (Chapters 3 & 4) explored the modulation effect in an active behavioral paradigm, where we probed infants' abilities to represent others' beliefs with a variety of contents. We explored whether

14-month-old infants can represent others' beliefs involving the individuation of multiple objects based on spatiotemporal or feature information. Crucially, we were interested whether those infants at this age who from their own perspective could vary the use of feature information for individuating purposes based on the validity of such cues; could also compute others' mental states relying on the same inferences; thus attributing to others false beliefs about identity. Thereby we aimed to provide an empirical investigation of some of the proposed limitations of infants' mindreading abilities (Butterfill & Apperly, 2013), and the sophistication of infants' ToM cognition. We then turned to probe other aspects of flexibility of such processes. In the final experimental chapter (Chapter 5) we explored how infants can integrate new information into existing representations of a goal-directed agent, and whether they can discard such representations when they face evidence suggesting their initial inferences were mistaken. To summarize, the present thesis aimed to address the following questions:

- Do signatures of representing primary representations also accompany representing them as belief contents? (Chapter 2)
- Does the modulation of one's own behavior by others' mental contents manifest itself in infants' active behavior? (Chapter 3 & 4)
- Can infants represent beliefs with a variety of contents that they themselves can entertain? Do infants show arbitrary limits in their ToM abilities, or do limitations correspond to their own representational abilities? (Chapter 3 & 4)
- Do infants represent mental states on-line, and can they flexibly modify them in light of new information? (Chapter 2 & 5)

Chapter 1

"Thoughts, perceptions are modes of subjective action— they are known only by internal consciousness and have no objective aspect. […] We cannot perceive the thought of another person at all, we can only infer it from his (its) behaviour."

- Charles Darwin (1840):
Old and useless notes about the moral sense & some metaphysical points

Humans are extremely social beings, possibly a lot more than even our closest primate relatives (Herrmann et al., 2007). As such, we need to be able to navigate in complex social situations to be able to interpret others' communication and behavior, to make inferences about them, and to use these to predict others' future actions. The most prominent signature of the human sociality is our ability to recognize that we ourselves, and others around us, have minds (Simon Baron-Cohen, 1995; Jacob, 2005)– and that the contents of the mind influence one's behavior. Humans generally represent states of their own mind (and can reflect on these), as well as ascribe mental states to others. This ability is usually termed as Theory of Mind (ToM) and includes reasoning about others' mental states such as beliefs, goals, or desires (Premack & Woodruff, 1978).

Theory of Mind as attribution of epistemic states – terminology and focus of the present thesis

The mental life of humans is rich, and mindreading encompasses a wide range of phenomena, therefore what should be considered as part of ToM is often unclear. Undoubtedly, many domains interact with each other and contribute to social cognition, therefore a full account of socio-cognitive capacities would require a great deal more than a doctoral thesis. We aim to discuss a subset of these capacities: the attribution of epistemic states to others by human infants; and to characterize the representational machinery and cognitive mechanisms that could enable it to take place. Before introducing the theoretical background, we aim to give a specification of the terminology and focus of the current work.

First, Theory of Mind has been termed in many different ways, and some have chosen one term over another to express their theoretical stance. For example, the theory-theory of ToM argues that our ability to represent others' mental states relies on a theory-like ability that is similar to scientific theories (Gopnik & Wellman, 1992). To distance himself from such theoretical stances, Ian Apperly in his book 'Mindreaders: The Cognitive Basis of "Theory of Mind" (Apperly, 2010) coins for the term 'mindreaders', to choose a term that is not theoretically laden. However, at the same time he points out that this term may reflect an analogy with reading insofar as reading entails interpreting written text, whereas mindreading entails interpreting observed behavior. More recently, Heyes and Frith (Heyes & Frith, 2014) argued for the acquisition of 'mindreading' to be similar to the learning of reading written texts. In the present thesis various terms will be used interchangeably, such as 'Theory of Mind' or 'ToM', 'mindreading', 'mentalizing', 'mental state attribution'; and we do not wish to commit to any implication of these terms unless specified otherwise. Most importantly, we endorse the term "Theory of Mind" because, as Premack and Woodruff (1978) described, it reflects the fact that mental states are not directly observable, and that the system can be used to make predictions, for instance about the behavior of organisms. The first point reflects one of the greatest challenges with regard to the studying of mental state ascription; and the second may be one of the main purposes of having such an ability.

Second, Theory of Mind can involve reasoning about a wide range of mental states that can be of different nature; such as epistemic (e.g. belief) motivational (e.g. desire), or emotional (e.g. scared). In most of this thesis we will focus on epistemic states, such as know or believe. We will therefore mostly discuss belief attribution

(Chapters 2-4), but we will discuss other kinds of mental states such as preferences and goals as well (Chapter 5).

Third, we are interested in representational mental states that take as their content specific states of affairs, therefore this thesis will mostly cover mental states that we refer to as 'episodic'. By episodic we mean that they are not beliefs like children's belief in Santa Claus, or some adults' beliefs in creationism; that are lasting and exert effect through a long period of time. Rather, they are beliefs that are linked to specific situations: they are formed based on objects and people (or other agents) that are present or otherwise perceivable, and based on events that involve the above; and are applicable within the frame of that situation. A typical situation is tracking the location of a particular object (or the lack thereof), which we will discuss later in this chapter. If, for example, after a dinner in a restaurant I see my friend giving her bank card to the waiter, but I don't see him bringing it back (as I happen to turn at that moment to find something in my bag), I will first correctly know that the waiter has the card, and then later mistakenly think that it is still with him. This belief, while it will probably influence my behavior in the restaurant in some way (e.g. I might not initiate that we leave, as I think we still need to get the card); it will unlikely have any long-lasting effects on my future behavior[1]. The latter, long-lasting beliefs are a topic on their own and not the focus of this thesis. We will discuss the process of attribution of beliefs that within a particular scene might influence an agent's behavior, such as their communicative or other actions.

[1] There can be exceptions to this, where e.g. my friend doesn't know about this year's vacation plans because she was not there on last year's vacation when we discussed this.

Lastly, the present thesis will discuss these phenomena primarily with the goal in mind to give a description of mentalizing capacities early in human development. As some of the theoretical questions arise, some aspects of the adult ToM literature will be discussed as well. However, there is a great amount of work out there that is relevant to the full picture of understanding the way living beings think about other minds. Hopefully in the future it will be possible to integrate knowledge from these subfields into a comprehensive theory of social cognition.

1.1. Theory of Mind: developmental findings and theories

In their seminal paper, Premack and Woodruff (1978) raised the possibility that chimpanzees impute mental states to others. They listed arguments that suggested (or would suggest, if tested empirically) that chimpanzees can solve tasks where they have to take into account someone's mental state as opposed to simply considering the physical constraints of a situation. Some 40 years later it is still an open question how much chimpanzees (Call & Tomasello, 2008) or primates in general (Martin & Santos, 2016) understand of others' mental states, but the article's largest influence came from a theoretical argument raised in the accompanying commentaries. Three philosophers (Bennett, 1978; Dennett, 1978; Harman, 1978) pointed out a challenge to the investigation of understanding other minds; namely that if an observer's mental state coincides with that of the other who she observes, we cannot know whether she attributes the mental state to the other, or simply predicts the other's action based on her own mental states. As Jacob points out, in this case *predicting* an agent's action would not involve *explaining* it, as for the latter one would need to represent the relevant psychological states of the agent (Jacob, 2013). The suggested

solution was to develop scenarios where the other's mental state is mistaken, as in this case the other's representation certainly must be different from one's own representation of the world (as one cannot represent something as reality that she does not hold to be true).

As a result, the most prominent tool used to investigate Theory of Mind abilities became the False Belief Task (FBT). In this task the crucial question is whether the subject can ascribe a different epistemic state than their own to another person, and use that to reason about the other person's behavior. Variations of this task have been used with typical and atypical human populations (Baron-Cohen, Leslie, & Frith, 1985) in several different cultures (Barrett et al., 2013) as well as other animal species (for a somewhat skeptical review see Penn & Povinelli, 2007). For the present purposes the most relevant aspect with regard to the development of Theory of Mind capacities is human children's performance on broadly two types of FBT. In particular, there is a well-known discrepancy between results on the elicited-response tasks and the spontaneous-response measures (terminology adapted from Baillargeon, Scott, & He, 2010). While children typically pass standard (elicited) false belief tests only around 4 years of age (Wellman, Cross, & Watson, 2001), in nonverbal paradigms they show sensitivity to others' beliefs[2] as early as 6-7 months of age (Kovács et al., 2010; Southgate & Vernetti, 2014).

[2] Some – maybe rightly – criticize the expression 'sensitivity' to others' beliefs (Csibra, personal comm.), as it is not clear what the 'sensors' would be in this case. But one could argue that just as the visual system is sensitive to light insofar as it processes specific kinds of input (in this case, waves of light) and provides an output from it (visual representations); the system responsible to represent mental states deals

In the next section we introduce the developmental findings and point out the challenges they raised. This will be followed by a critical discussion of the theoretical approaches of ToM development. We will argue that accounts positing a radical conceptual change from infancy to childhood face both theoretical and empirical challenges; and will suggest a model of early ToM abilities that is congruent with the gradual development of the domain. We close the introduction with outlining some of the predictions of this suggested account, as well as highlighting open questions with regard to infants' ToM capacities that are the focus of the current work.

1.1.1. Assessing ToM development: empirical findings

Findings from elicited tasks

In its original form of the False Belief Task, which is recently referred to as elicited-response FBT (Baillargeon et al., 2010), subjects are asked to predict the protagonist's action based on her mistaken belief. In order to respond correctly, subjects need to take into account the knowledge state of the other, that differs from their own knowledge (and from the true state of affairs). For example, in the change-of-location variant of the task (Baron-Cohen et al., 1985; Wimmer & Perner, 1983), a protagonist, say Sally, first puts her toy into a box, and then she leaves the room. In Sally's absence, her friend Anne takes out the toy from the box and puts it into a

with specific kinds of input (e.g. information on agents and their potential perceptual access), and provides output (e.g. mental state representations). Therefore by saying this we simply mean that infants possess the necessary cognitive apparatus to process information that is relevant for e.g. representing someone's belief.

basket. Participants are asked to predict where Sally will look for her toy. To answer correctly they need to inhibit the actual location of the toy, and instead give the incorrect location as the answer. Other variations of this task include giving similar verbal answers based on a protagonist's false belief about the content (Hogrefe, Wimmer, & Perner, 1986) or about the identity (Apperly & Robinson, 1998) of an object.

A comprehensive review by Wellman and colleagues (Wellman et al., 2001) found that children tend to pass such tasks around the age of 4, with younger children showing systematic failures involving giving the actual location of the toy as an answer. This pattern was found to be relatively consistent across different tasks (Gopnik & Astington, 1988; Perner et al., 1987), with children's performance on variations of the task with regard to belief content (location vs. identity of the object) showing a correlation with each other (Rakoczy, Bergfeld, Schwarz, & Fizke, 2015). While some have argued for a universal, synchronized onset of the capacities allowing to pass ToM tasks (Callaghan et al., 2005) others confirmed a similar trajectory of development but with cross-cultural variations in the age to pass particular tasks (Liu, Wellman, Tardif, & Sabbagh, 2008), or the order in which different types of tasks are passed by children (Shahaeian, Peterson, Slaughter, & Wellman, 2011).

Successful performance on these tasks has been declared to signal a fundamental conceptual change (Perner et al., 1987; Rakoczy et al., 2015), as it was argued to show that children have acquired a representational Theory of Mind (Gopnik & Astington, 1988; Leslie, 2000a); that is, the understanding that beliefs are in fact

representations[3]. This view has been heavily criticized from many directions. For instance, on theoretical grounds Leslie argued that infants in fact start out with having a 'belief concept', and "It is the possession of the concept BELIEF (plus a gradual increase in skill at employing the concept) that eventually gives rise to a commonsense representational theory of mind." (Leslie, 2000, p.14). Others have challenged the use of success on false belief tasks as a litmus test for 'having a Theory of Mind', as it poses unnecessary challenges on children, therefore not passing such a task is not informative with regard to a child's conceptual abilities (Bloom & German, 2000). In line with this, Zaitchik (1990) found that those children who fail the false belief task also fail the false photograph task[4], whereas studies involving older children with autism showed that they tend to pass the false photograph (but not the false belief) task (Leslie & Thaiss, 1992); which suggests that failure on the false belief task can be due to different reasons.[5] Even if children pass, it might not reflect full

[3] Some theorists (Gopnik & Wellman, 1992) claim that before age 4 children have a 'copy theory of belief'; that is, think that others' representations are simply copies of current states of affairs, and only at age 4 do they begin to understand that people's representations may be incongruent with reality. We will discuss different approaches regarding children's abilities before age 4 in section 1.2.2.

[4] In this task it is not a person's belief that misrepresents the current state of affairs, but a photograph that was taken before a change has happened, which the child observes.

[5] The false photograph was later criticized to not be analogous to the false belief task, as photographs are not false but rather outdated. As a response a better-matching false sign task was developed; as signals serve the purpose of correctly representing a current state of affairs in the world. On this task, again young children's performance

understanding of mental states, as it might be solved through alternative strategies (Fabricius, Boyer, Weimer, & Carroll, 2010). Finally, others pointed out that there is still improvement on more complex ToM tasks after the age of 4 (e.g. Apperly & Robinson, 1998; Simon Baron-Cohen, O'Riordan, Stone, Jones, & Plaisted, 1999). Supporting evidence comes from neuroimaging data showing that the specificity of activation in some of the brain regions that are argued to be related to ToM reasoning, such as the temporo-parietal junction or TPJ (Frith & Frith, 2003; Saxe & Kanwisher, 2003) still seem to undergo some changes between ages 5 to 11 (Gweon, Dodell-Feder, Bedny, & Saxe, 2012).

In sum, while there is considerable debate on the role and significance of elicited false belief tasks, and many modifications on the task have proven to be informative about the underlying capacities (e.g. Scott & Roby, 2015; Setoh, Scott, & Baillargeon, 2011), the basic finding remains that even if specific aids are implemented to help the child through the task (e.g. Rubio-Fernández & Geurts, 2013), the youngest age to succeed is around 3 years.

was found to be correlated with the false belief tasks, and contrary to the false photograph task, older children with autism showed poor performance, even on nonverbal versions (Leekam, 2016). Nevertheless, these results don't exclude the possibility that in elicited false belief tasks domain-general task demands interact with children's competence.

Findings from spontaneous measures

Evidence supporting belief understanding in infancy comes from spontaneous measures, where the participant is not asked to reflect on someone's mental state, but they are exposed to belief-involving scenarios, and their spontaneous response is assessed. In a variation of the location-change false belief task (Clements & Perner, 1994; Garnham & Ruffman, 2001) children between 2 and 3 years of age were found to look correctly where a protagonist should appear based on her belief; suggesting that if we measure children's spontaneous responses (e.g., their looking patterns) instead of asking them direct questions about a protagonist's beliefs and her consequent behavior, even younger infants may show sensitivity to others' mental states. Indeed, in their influential violation-of-expectation study Onishi and Baillargeon (2005) found that infants at 15 months were found to expect a protagonist to search for an object at the location where she (falsely) believes it to be hidden, which was suggested by their looking times during the different scenarios. Specifically, infants looked longer if the agent acted as if she had a true belief, while in fact she was mistaken about the location of an object; suggesting that this event was unexpected to them.

Further studies have shown that at 2 years of age infants predictively look to the location where an agent will search based on her false belief (Southgate, Senju, & Csibra, 2007; for non-human agents see Surian & Geraci, 2012) that are based on her perceptual access (Senju, Southgate, Snape, Leonard, & Csibra, 2011), around one year of age they look longer if an agent acts inconsistently with her beliefs or perception (Song & Baillargeon, 2008; Surian, Caldi, & Sperber, 2007), and even at 7 months their reactions are influenced by another agent's beliefs (Kovács et al., 2010).

Recently, Southgate and Vernetti (2014) provided converging electrophysiological evidence showing that 6-month-old infants make action predictions (sensorimotor alpha suppression signaling motor cortex activation) based on an agent's belief. Such results were argued to indicate that young infants represent the other agent's mental states (most typically beliefs, but see Luo, 2011; for preference attribution based on the person's belief in 10-month-olds). Additionally, infants in their first year of life can interpret the actions of agents as goal-directed (Csibra, 2008; Gergely & Csibra, 2003; Luo & Baillargeon, 2005; Woodward, 1998) which has been linked to later (explicit) Theory of Mind abilities (Aschersleben, Hofer, & Jovanovic, 2008; Sodian et al., 2016; and for relationship between understanding intention and later ToM abilities see Wellman, Phillips, Dunphy-Lelii, & LaLonde, 2004).

Infants not only show an understanding of others' beliefs in relatively passive contexts, but they also incorporate these into their behavior and react to initiations of interactions accordingly. There is a set of evidence showing that 12-month-olds are able to correctly infer the referential intention of an adult. For instance, they assume that an adult is likely to point at (and request) an object she has previously interacted with (Liszkowski, Carpenter, Striano, & Tomasello, 2006). Other studies suggest that 17-month-olds respond to object requests according to the person's belief: they retrieve the object from the correct, rather than the pointed-to location when the agent has a false belief about the location of the object (Southgate, Chevallier, & Csibra, 2010). Similarly, if a person is looking for an object she has put in one of two boxes, 18-month-olds respond to this person's request for help based on what she believes about the boxes; if she has not seen the location swap they open the box that actually contains the object, rather the one she tries to open (Buttelmann, Carpenter, & Tomasello, 2009). Such helping behavior dependent on the protagonist's belief was

observed in an unexpected-identity task as well (Buttelmann, Suhrke, & Buttelmann, 2015). Furthermore, 18- and 24 month-old infants anticipate another person's actions based on her false beliefs and point to correct her belief even *before* she initiates her action (Knudsen & Liszkowski, 2012b).

Thus, infants have been shown in a variety of scenarios, with a wide range of methodology, to display sensitivity to other agents' beliefs. Crucially, all of these studies involve spontaneous measures, where subjects were not directly asked to reflect on someone's mental states. These findings challenge the view that ToM abilities necessarily emerge during the preschool years, and provide promising evidence for theorists claiming an early origin of mental state understanding. In the next section the various theories are described, with a focus on their interpretation of developmental findings.

1.1.2. Explaining ToM development: theoretical approaches

Any theory on the development of Theory of Mind needs to address the question how findings from elicited tasks showing ToM competence emerging around 3 years of age relate to those from spontaneous-measure paradigms that go as young as 6-7 months. The various approaches differ with regard to how much they assume that spontaneous tasks measure competences related to ToM; and if they do, whether that is something fundamentally different from the fully developed mindreading capacities. The gradual change accounts posit an early emergence of ToM, and point out that elicited tasks require abilities that are unrelated to ToM abilities, and that are still under development in preschool years. The conceptual change accounts claim

that infants' success on the spontaneous tasks does not reflect ToM capacities, but infants' behavior can be explained by some simpler cognitive mechanisms. Two-system accounts mostly agree that infants' behavior reflects some form of mindreading (or its precursor), but they differ in what ability they grant infants and how they see subsequent development.

Gradual development accounts

Some theories posit that the core mechanisms of ToM abilities are present from birth, and while they get enriched through development, this does not involve the acquisition of radically new concepts, but rather gradual sophistication of both ToM and various other domains. One alternative was suggested by Baillargeon and colleagues (Baillargeon et al., 2010), who propose two separate subsystems for belief reasoning, developing sequentially. One (subsystem-1) deals with reality-congruent states, such as and goals and dispositions, and is present by the end of the first year of life. The other (subsystem-2) deals with reality-incongruent informational states, such as false beliefs, and starts to operate in the second year of life.

An earlier account comes from Leslie (1994), who argues for an innate ToM module. According to his view, TOMM (Theory of Mind Mechanism) has a first version, TOMM1, which ensures the understanding of goal-directed actions. The second, TOMM2, makes the understanding of beliefs and desires possible. Later the theory was extended with the SP (Selection Processor), which enables explicit reasoning about beliefs and dealing with the inhibitory demands of the task. In fact, it is the lack of maturity of SP that is argued to be responsible for preschoolers' failure on elicited ToM tasks. There is a suggested stepwise development in this theory, but

at the same time it assumes that metarepresentational capacities, together with the ability to represent propositional attitudes[6], should be present from birth (Leslie, 1987). Both Leslie's account, and that of Baillargeon's (Baillargeon et al., 2010) describe two systems *within* belief reasoning, therefore can be rather considered as multi-componential (Carruthers, 2015b). The idea to consider ToM as constituting of various sub-components has been suggested by others as well; the common suggestion being that rather than positing one or more large (modular) systems responsible for the entire process of belief reasoning, to view it as an orchestrated functioning of many different processes, only one of which is verbal reporting of previously calculated belief representations (Christensen & Michael, 2016; Kampis, Fogd, & Kovács, in press; Kovács, in prep.).

Should one grant infants ToM abilities of some extent, then it bears an explanation why these do not manifest themselves in children's responses on the elicited tasks until a certain age. According to the response or processing-load account (Baillargeon et al., 2010) the ability to represent false beliefs indeed emerges early, but children fail the elicited-response tasks until a certain age because these tasks require several other capacities; such as the selection of the correct response (through accessing their representation of the protagonist's belief); and inhibition of other responses, such as the prepotent response that is based on their own knowledge (Leslie & Polizzi, 1998; Leslie, German, & Polizzi, 2005). In accordance with this view, some studies have found correlation between inhibitory control and

[6] A propositional attitude is described as a "computational relation between an organism and a mental representation expressing the proposition *p*. " (Leslie, 2000a, p. 10)

ToM abilities (Carlson, Moses, & Claxton, 2004), and more advanced ToM in bilinguals (Kovács, 2009) who are known to have advanced executive functioning (Bialystok, Craik, & Luk, 2012).

In addition to correlational data, some have found better performance if processing load is lowered, for instance the object had been removed and hence the salience of the true response did not have to be inhibited (e.g. Koos, Gergely, Csibra, & Biro, 1997; for a summary of similar studies see Wang & Leslie, 2016, and for counterevidence see Call & Tomasello, 1999), or other demands of the task have made it easier to follow (Rubio-Fernández & Geurts, 2013); and worse performance if processing load is increased in an otherwise matching task (Scott & Roby, 2015). In line with this, there seems to be a moderate, but relatively consistent correlation between linguistic abilities and ToM (Milligan, Astington, & Dack, 2007). Additional evidence comes from training and longitudinal studies with children that have found a relationship between elicited and spontaneous measures, suggesting that the two build on the same underlying abilities. For instance, longitudinal data suggests a relationship between early spontaneous ToM and later performance on elicited tasks, although only if the tasks match in terms of content (Thoermer et al., 2012). Relatedly, Clements and colleagues (Clements, Rustin, & McCallum, 2000) found that a training for preschoolers on false belief was only effective in the group of children who showed an implicit understanding before the training.

While the external task demands of the elicited ToM tasks are not debated, recently, the performance account has been criticized for falling short of explaining the full picture of ToM development. First, as Carruthers (2013) points out, some of the tasks that infants *do* pass, are nevertheless likely to rely on executive functioning,

such as helping tasks (Buttelmann et al., 2009). Moreover, there is some evidence that performance even on implicit tasks is correlated with EF measures in 18-month-olds (Yott & Poulin-Dubois, 2012), and increasing EF demands in adults can interfere with implicit belief processing (Schneider, Lam, Bayliss, & Dux, 2012). While performance accounts indeed do not address such effects, it is possible that extraneous factors such as executive function processes play a role in both elicited and in spontaneous tasks. In addition to these task demands, however, in elicited verbal tasks there is an extra burden that falls onto the child: she needs to give an appropriate answer that expresses well the mental state of the agent she is asked about (Carruthers, 2013).

Other accounts focus on the pragmatic aspects of mindreading. On one hand, pragmatics play a role in elicited tasks themselves, as supported by studies showing that modulation of task questions influences children's performance (Clements & Perner, 1994; Garnham & Ruffman, 2001; Helming et al., 2014; Siegal & Beattie, 1991). A very intriguing recent study found converging evidence that pragmatic factors play a role in performance on the True Belief questions as well (Oktay-Gür & Rakoczy, 2016). With a subtle manipulation of the standard elicited location-change task the authors targeted the question why children tend to fail True Belief questions, especially after they pass False Belief ones. They argued that pragmatically it might be confusing to the children if they are asked about something that in fact everyone who is present, knows. Consistently with this hypothesis, they found that introducing two protagonists; one with a true and one with a false belief, helped children answer correctly both questions. Creating a contrast between someone who *doesn't* know, and someone who *knows* the correct location, made it pragmatically justified asking about the true belief; even though with regard to memory load the two-agent condition should have been more taxing for children.

On the other hand, Westra (2016) argues that pragmatic development has a more substantial role in passing the verbal-elicited FB tasks. According to this proposal even though infants and children might have the necessary underlying conceptual apparatus to represent agents and mental states; they would need to learn when it is relevant to talk about them in discourse. This account works with the assumption that ToM has innate origins, and outlines a possible mechanism how the early capacities may develop into more mature mindreading[7].

The accounts discussed in the present section suggest that ToM abilities are present from early on, and go through gradual development to preschool age and further. The present thesis aims to characterize some of the cognitive mechanisms that may support early mindreading capacities. Before turning to the characterization of these processes, we discuss approaches that posit systematic limitations on early ToM abilities and argue for a radical conceptual distinction between earlier abilities and later developing, full-blown Theory of Mind.

[7] One challenge for this account is how mental state representations become accessible for report in the first place, as incorporating them into discourse requires that they can be reported on.

Conceptual change accounts

Conceptual change theories share the assumption that infants' cognitive abilities should be distinguished from mature ToM abilities, because the conceptual apparatus required for representing mental states is not present at a young age. These accounts explain infants' behavior with mechanisms that are external to mindreading. One proposal is that infants succeed on spontaneous tasks by using behavioral rules (e.g. 'people tend to look for an object where they last saw them'), or three-way associations between the actor, the object and the location (Perner & Ruffman, 2005). Contrary to the latter (three-way associations) hypothesis, Baillargeon and colleagues (2010) point out that in many cases there is a difference in looking time in one condition, while it is absent in another similar condition where the same associations would be possible (e.g. Luo & Baillargeon, 2005, 2007). Against the former, behavior-rule interpretation the body of experiments grows where infants in different situations act as if they take into account the protagonist's false beliefs, which makes this account decreasingly plausible, in addition to being post-hoc in describing what rules should be at play (Carruthers, 2013). First, infants show similar pattern of reaction in situations they probably have not encountered before and thus they could not have formed a corresponding rule (such as actions of squares). Second, as Surian et al. (2007) point out, some rules are unlikely to be learnt (and even less likely to be innate). Third, more and more rules would have to be implemented; therefore it becomes less and less parsimonious to assume acquired behavioral rules (Baillargeon et al., 2010). Further evidence against this account comes from the study of Southgate and colleagues (Southgate et al., 2007), where they rule out the application of many of the low-level behavioral rules, such as the infants simply orienting to the first or the last place the ball was seen before, or the most recent place the actor looked at.

There have been accounts suggesting that in children attributing ignorance may be present earlier than understanding (false) beliefs (Hogrefe et al., 1986). However, others excluded the possibility that attributing ignorance may explain infants' behavior, suggesting that infants treat someone with a false belief different as an ignorant actor (Knudsen & Liszkowski, 2011). Moreover, as Martin and Santos (2016) pointed out, representing ignorance also requires to attribute a representation that is decoupled from reality, therefore "representing agents as truly ignorant likely requires the same cognitive resources as representing agents with false beliefs, namely, forming a representational relation" (Martin & Santos, 2016, p. 379).

Finally, according to a more recent account by Heyes (2014a, 2014b) most effects that are found in infants are claimed to be due to low-level, domain-general processes. Specifically, it argues that perceptual and imaginal novelty, that are driven by the saliency of the stimulus and memory processes such as retroactive interference, would be responsible for infants' patterns of surprises. In a detailed description of some of the infant paradigms Heyes (2014a) claims to account for all infant findings through an alternative description involving the above processes, where in each study a different combination of them would be responsible for the observed looking patterns. This approach, while seemingly thorough (by listing several experiments in detail), is highly unlikely to be correct, due to several considerations. First, as in their reply Scott and Baillargeon (2014) point out, some of the alleged underlying mechanisms seem to be exaggerated or unfounded. For example, Heyes (2014a) reports a "delay-related memory limitation", where during subsequent trials the memory of the previous ones would completely fade; but in fact there are no good grounds to believe that such an effect exists (and the account

doesn't provide such grounds). In fact, a recent study looking at a possible connection between implicit false belief understanding and susceptibility to memory interference and distraction, found no relationship between these two (Zmyj, Prinz, & Daum, 2015). Another line of arguments against this approach comes form the fact that the applications of the rules in specific studies seem arbitrary and post-hoc. It is often unclear whether beforehand similar predictions should be made and on what grounds; and if in some studies instead of one mechanism (e.g. retroactive interference) another one (e.g. saliency) was applied, it could reverse the predictions. In sum, while the general approach of scientific scrutiny is one to follow, and 'rich' interpretations should be handled with caution; Heyes' account does not seem to provide a coherent picture of the socio-cognitive abilities of infants.

Even if one does not grant infants any mindreading abilities, all (or at least those who believe adults are capable of such cognition) accounts agree that during the preschool years children demonstrate an understanding of the representational mind. Competence accounts argue that such abilities emerge due to some maturation, or cultural learning. To date, no account exists that could give a full explanation of how the foundations of these capacities develop, should they not have some rudimentary innate basis. For instance, recently Heyes and Frith (Heyes & Frith, 2014) took the term mindreading in a much too literal sense, and argued that there is a parallel between how children learn to read books, and to reason about others' mental states. However, they do not provide a mechanism how mental state concepts should be learnt by someone who does not already possess them. Moreover, some of the mechanisms (such as communication, teaching) that they suggest to be the source of acquiring a ToM, themselves arguably require some understanding of (e.g. the communicator's, or the teacher's) mental states (Jacob, 2014).

Mental file theory

Perner's account on the other hand, makes very specific suggestions about the cognitive changes that occur around age 4, and enable the emergence of representational ToM. This account will be spelled out here with some detail, as it gives fruitful ground to the description of cognitive mechanisms underlying ToM. First, Perner claims that representing beliefs entails metarepresentations; for which he uses the definition of Pylyshyn, namely representing representations *as* representations (Pylyshyn, 1978).[8] By this he also refers to the notion of 'intensionality' or 'aspectuality' of beliefs, namely that "belief about an object depends on the label under which the object is known to the believer" (Perner, Huemer, & Leahy, 2015, p. 78). In this description, Perner refers to Frege's distinction between 'sense' and 'reference', and claims that understanding this distinction is crucial for children to represent beliefs (Perner et al., 2011). In this terminology, the 'referent' of a statement is the external entity about which the statement is made; for example for the statement "the yellow key is on the table" has an external referent; an actual key, that is physically present on the table. However, the statement also has a 'sense', or a mode of presentation or aspect under which this key is represented; namely the aspect of it being a *yellow* key. However, it is possible that we can also say of the same key: "the green key is on the table". This statement might refer to the same external referent, the key; but will have a different sense. The two statements can be hence connected by an identity statement "the yellow key is the green key", which

[8] The validity of this definition itself is debated, as we will return to it later in the introduction.

statement is only informative if one is sensitive to the different senses under which the key can be represented. In this case, the identity statement will inform that the two senses belong to the same external referent.

Perner argues that to grasp the notion of beliefs, one needs to understand the aspectuality of belief representations. Namely, that someone might have a particular description (or 'sense') of an external referent that can differ from one's own sense of representing the same referent[9]. On this account, children are unable to grasp the aspectuality of beliefs before the age of 4, which comes from the characteristics of how children handle 'mental file' representations (outlined in more detail below), both from first- and from third-person perspective.[10] In accordance with this claim, they found a correlation between identity statements and performance on the verbal false belief task (Perner et al., 2011). Curiously, initial findings showed a delay in the understanding of identity-related false beliefs (Apperly & Robinson, 1998). Four to six-year-old children were told a story involving two objects: a regular eraser and one that was also a die. Children were asked whether a puppet, who was not informed about the dual nature of the die-eraser (but would simply perceive it as a die), would know that the die is also an eraser. Children correctly answered 'no'. However, then

[9] Flavell's (Flavell et al., 1981) distinction between Level-1 perspective taking (*whether* something is seen by another person), and Level-2 perspective taking (*how* something is seen by someone) is analogous to this: in Level-2 perspective taking one referent can come under two different descriptions (one's own, and the other's).

[10] Other accounts also argue that infants fall short of understanding the aspectuality of beliefs (Apperly & Butterfill, 2009; Butterfill & Apperly, 2013), we will discuss these in the next section.

they were asked where the puppet would look for an eraser, they incorrectly chose at random between the two objects. Apperly and Robinson (1998, 2001) concluded that children at this age have difficulties with situations that require understanding intensionality. However, in a simplified version of this task, involving just one object, Rakoczy and colleagues (Rakoczy et al., 2015) found that children pass such tasks around the same age as location-change tasks, and children's performance on the two kinds of tasks correlated with each other. While the Rakoczy et al. findings support Perner's claims through suggesting that when children are capable of metarepresentational reasoning about beliefs, they can do so with different kinds of belief contents; Perner (Perner et al., 2015) highlights a curious effect that was found in a modified version of this task (Sprung, Perner, & Mitchell, 2007). Here, children answered action-prediction questions correctly when they were not asked in an 'individuating manner' (e.g. a die or an eraser) as in the previous studies, but rather in a 'predicative manner' (e.g. a long stick vs. a short stick).

The distinction between individuating and predicative information plays a central role in Perner's mental file theory (Perner et al., 2015; Perner & Leahy, 2016), developed based on Recanati's notion of mental file (Recanati, 2012). Here, the main assumption is that mental files are tools that capture information about objects in the world – about *referents* (cf. Perner et al., 2011, as described above). Every file is anchored to an object that it contains information about; individuating information fixes the referent, and predicative information provides information about the referent on the file. Files can capture different perspectives on the same object: both 'conceptual' perspectives (e.g. die or eraser) and 'mental' perspectives (how different people perceive the same object). For each perspective on the same object a new file is opened, and through this each file individuates the object in one specific way. Then

in each file there is predicative information, e.g. 'the die is in the box'. In addition, there need to be links between the files, in order to capture that different files might be anchored to the same external referent; and identity statements are precisely serving this function, namely to link together different files belonging to the same object. Through this mechanism the intensionality/aspectuality problem is grasped: one can represent the same extensional object through different intensional descriptions, and this would be operationalized through having separate files for the same object, each file grasping one intensional description. Crucially, besides one's own 'regular' files, there are also 'vicarious' files that are files attributed to others. Similarly how so-called horizontal links are used to link one's own mental files; vertical links are used to connect regular files with vicarious files, in order to establish the common external referent.

According to the mental file theory, it is the capacity to form such horizontal and vertical links that develops around the age of 4 years. It is responsible for children's understanding of identity statements, as well as children's metarepresentational understanding of other's beliefs. The mental file theory also offers a possible explanation for infant data; that infants would not individuate files by linguistic terms, but they may use non-linguistic conceptual information. This way different kind-properties would be stored as predicative information within the same file; and hence the problem of linking would not occur. However, already young infants use labels as a basis for individuation (Xu, 2002), moreover, they also use other nonverbal information such as functions for kind-based individuation (Futó, Téglás, Csibra, & Gergely, 2010); therefore the conjecture that they would not use these to open separate files, does not seem to hold.

An additional suggestion for explaining infant data comes from Perner (2016), that vicarious files would be represented as just another regular file, and not linked to one's regular file. Simply representing vicarious files would not be metarepresentational, but rather simulative; as metarepresentation entails linking the vicarious files to regular files, and to represent this link between the two. This linking capacity emerges at age four, before that infants represent vicarious files 'without knowing so'; and whether they can switch to the vicarious file from their own file depends on the saliency of external cues. Once they become able to form this link, intentional switch to it becomes possible. While mental files may indeed be a useful tool to capture how infants store representations about the environment or about others' beliefs; Perner's explanation of how infants handle vicarious files faces a challenge, when applied to false belief scenarios. How would infants know, when to access the vicarious file (e.g. to update the predicative information on it), and how would they access the vicarious file at all? Perner's suggestion is that these unlinked vicarious files are accessed when a file-related event happens; but this suggestion seems rather arbitrary; as it is unclear how one would know what are the file-related events without knowing that the file 'belongs' to someone else. In fact, with regard to identity statements from first-person perspective there is only evidence on *verbal* tasks suggesting that children have problems (and alternative naming tasks; Perner, Stummer, Sprung, & Doherty, 2002), that would support the claim that children (and infants) do not have the capacity to link files to each other. As linguistic tasks have been argued to pose challenges to children, and to our knowledge to date there is no task that measures such capacities nonverbally; it could well be that in fact the capacity to *link* files is present from as early as the capacity to *form* mental files. If linking files is functional already in infants, then they may represent beliefs via

mental files just like preschoolers or adults. We will build on this possibility in Chapter 1.2.1.

Some findings suggesting that the ability to link files is present from early on come from Cacchione and colleagues (Cacchione, Schaub, & Rakoczy, 2013), who familiarized one group of 14-month-old infants with regular objects, and another group of infants with transformable objects. Then in test, infants observed object A being put in an opaque box; and object B retrieved. In fact, A and B were one and the same object, but observing this event should result in the individuation of two different objects; and hence the (mistaken) opening of two files. Indeed, the group who was familiarized with regular objects, acted after the retrieval of B as if they assumed the original object (object A) still be in the box, and continued to search for it. However, the group who was familiarized to transformable objects, searched *less*; compatible with the notion that they *linked* the two files they opened for the two object appearances (A and B), and inferred that therefore no object has remained to search for. While it is possible that this latter group did not successfully link the two files but rather did not open two files for the two appearances at all; this is quite unlikely, as it would mean that by a short familiarization infants have given up their expectations that different appearance signals different objects; and would expect *any* object to be the same as *any other*. It is more likely that in nonverbal cases, when the linguistic and pragmatic demands are lower, even young infants are able to make the necessary links between regular files. This alternative leaves open the possibility that infants are able to represent and link vicarious files as well, and hence early mindreading abilities share their core mechanisms with those of older children.

In sum, Perner's mental file theory gives a detailed description of the mechanism to represent beliefs. While it does not grant infants the full ability to operate on such representations, it currently does not offer any strong evidence against this possibility. In fact, the possibility that mental files could serve as the content of belief representations might be a useful starting point to describe mindreading abilities from early on.

A two-system approach: the minimal-ToM proposal

The 'minimal ToM' account of Apperly and Butterfill is an approach that suggests two separate systems for mindreading (Butterfill & Apperly, 2013; Low, Apperly, Butterfill, & Rakoczy, 2016; Rakoczy et al., 2015). This proposal starts with the assumption that representing others' mental states via propositional attitudes is resource demanding; because they can have embedded contents, and can interact with each other in complex ways, handling them should pose demands on working memory, executive function, and attention. However, often one needs to act fast, but still act according to the mental states of others. Thus as a solution to this proposed tradeoff between efficiency and flexibility of mindreading, the two suggested systems can be considered complementary; which they describe to be similar to the different systems involved in number cognition. There, it was suggested that the early developing systems have signature limitations, and these early systems are developing into, or are combined by, a later developing, flexible system that can overcome these limits (Carey, 2011; Wynn, 1992). Drawing on this analogy, System 1 in Apperly and Butterfill's account is fast and efficient, but is limited in its functions; and System 2 operates in a more resource demanding way, but is capable of more complex computations. As such, System 2 is referred to as full-blown ToM, and it has

no restrictions in the scope of representations it can handle. It represents propositional attitudes, and is involved in verbal reasoning about mental states (and therefore in all elicited/verbal tasks). However, it is not fast enough to be used spontaneously in online social interactions[11].

System 1 operates in a way that closely resembles mindreading, but is said not to be mindreading 'proper'; it is fast and efficient, but it has limitations that come from the format of representations it uses. This system is present from early infancy (and potentially exists in nonhuman animals), and it does not use propositional attitude representations. Rather, it involves computing representations that are called 'registrations'. The first basic principle is that infants can represent goal-directedness in a non-mentalistic way, rather as actions bringing about (or be directed at) goals. What determines what an agent can act on in a goal-directed way? This system can compute whether an agent has 'encountered' an object (which is a proxy to the notion 'perceive'); namely, objects that are within her 'field' (which is a set of objects that are within physical proximity of the agent; a proxy to what she would have perceptual access to). An agent can only act on an object in a goal-directed way, if she has encountered it. These encounterings are relations, and not representations; therefore representing them does not require metarepresentations. Moreover, an agent can only successfully act on an object in a goal-directed way if the object is still there

[11] While the minimal-ToM account proposes parallels with the cognitive systems involved in numerical cognition, analogies can also be found with dual-system proposals of reasoning, where System 1 is argued to be innate, rapid, and automatic; and System 2 is slow but capable of abstract thinking not permitted by System 1 (for a review see Evans, 2003).

where she encountered it. This is captured by the concept of 'registrations', that are relations between an agent, an object, and a location. An agent registers an object at a particular location if that is where she most recently encountered it; this should ensure that registrations, similarly to beliefs, can be correct or incorrect. Therefore, successful goal-directed action will depend on the agent having a *correct* registration of the object. Furthermore, an agent will always act according to her registration; therefore if the registration is incorrect, her action will lead to some sort of failure. This latter description can function as a proxy for false belief cases, and is argued to explain the majority of ToM studies that were conducted with infants. In addition, the authors state that features of this system could be extended in a way that is also incorporates other mental states, such as desires (Butterfill & Apperly, 2013).

Since registrations (involved in System 1) are relations between an agent, an object, and a location; the 'minimal ToM' account by Apperly and Butterfill (Butterfill & Apperly, 2013; Rakoczy et al., 2015) circumvents the problem of propositional attitude representations (and therefore metarepresentations), which are suggested to be costly to represent. The theory also makes predictions about the limits on the kinds of representations System 1 can handle; analogously to the "signature limits" of the number system, such as the 3-item limit of the system that tracks exact numerosity (Feigenson, Carey, & Hauser, 2002). Specifically, registrations are agents' relations to *objects*, and not to *representations of objects*, therefore it should not be able to handle mental states involving quantification or identity. This is because representing these would involve agents' relations to *representations* (cf. Perner's *sense*); as neither can be captured by a relation to one or more *actual* objects. Ascribing to an agent, say, a belief about which location has the 'most' objects, would not be possible through encoding the agent's relations to objects, but would require

representing the agent's attitude towards the abstract representation of the quantifier 'most'. Similarly, if one could only represent an agent's relation to an object itself, handling incorrect registration about object identity (*how* an object is registered), or about the absence of an object (to represent that an object is *not* registered) would be not possible. As according to proponents of this view current data provides no satisfactory evidence that these capacities are present in infants, one goal of the present thesis (in Chapters 3 & 4) is to take this challenge and probe infants' understanding of various belief contents.

The distinction between the two mindreading systems maps onto the distinction between Level-1 and Level-2 perspective taking (Flavell, Everett, Croft, & Flavell, 1981). Level-1 perspective taking involves judging *whether* a person sees for instance an object. Such computations could be performed by System 1 described above. However, Level-2 perspective taking could be described as representing *how* a person perceives an object; which is in fact involved in representing beliefs about identity of objects. Therefore Level-2 perspective taking has been argued not to be possible to perform with System 1. As a consequence, there should be no spontaneous Level-2 perspective taking, and no spontaneous tracking of beliefs about identity, which is congruent with some findings (Low & Watts, 2013; Surtees, Samson, & Apperly, 2016), but as will be discussed later, has been recently challenged by other studies (Elekes et al., 2016; Surtees, Apperly, & Samson, 2016). As representing false beliefs about identity is a central issue in the debate on the format of early belief representations, we will return to this claim later.

Finally, the minimalist account is somewhat ambivalent with regard to the relation between the two systems. While they do not exclude the possibility that the

two systems are somehow related, they state that the two cannot be fully continuous; that is, System 2 cannot be simply the explicit counterpart of System 1. Most importantly because of the limitations of System 1 in what kind of belief contents it can represent: some contents, such as beliefs about identity, could only be represented in System 2. In addition, the proposal of two systems was based on a speed-flexibility tradeoff, therefore they should not be able to handle the same scope of situations. Therefore, they take it as a viable option that the two systems work completely independently; but even if they might operate in parallel, they are considered to be two truly distinct cognitive systems.

Critiques of the minimalist account

While the 'minimal ToM' approach answers some of the issues, it also raises new ones. For instance, as mentioned above, the theory is vague on the relation between the two systems. It seems to be problematic and quite unlikely, that there would be two ToM systems that do not communicate with each other, and without any precursors the second system pops up at the age of four (De Bruin & Newen, 2012). While proponents of this view draw on a parallel with the number systems; this analogy does not completely hold. In the case of numbers, one leading idea is that the system that represents small numbers serves the primary function of tracking objects (Feigenson et al., 2002), and only implicitly represents number, through the number of objects tracked. A similar justification does not exist in the case of the two mindreading systems: they are two systems that aim at representing the same phenomena, and end up representing it in two different ways. Nevertheless, in the case of numbers; while there is no clear consensus of how various systems interact, to our knowledge it is not assumed that there should not be any connection between the

two. In fact, one of the challenges still is to explain how the development and operation of the two would be orchestrated, with some promising accounts emerging (Carey, 2011). Analogously, any account that supposes two distinct ToM systems, should have a story on how they are related to each other, both developmentally[12], and functionally (Christensen & Michael, 2016). While in the case of number cognition the goal of learning is to combine the two systems, in the case of Apperly and Butterfill's view it is unclear what role learning would play in ToM development.

In fact, there is evidence indicating that implicit and explicit mentalizing rely on the same core mechanisms. Recent neuroimaging data suggests that implicit and explicit inferences about other people's traits activate the same mentalizing areas, and ERP studies reveal that goal and trait inferences triggered by implicit and explicit instructions have a similar early timing (for a review see Van Overwalle & Vandekerckhove, 2013). With regard to inferences in belief reasoning, neuroimaging data (involving fMRI: Kovács et al., 2014; or NIRS: Hyde, Aparicio Betancourt, & Simon, 2015) suggest that the temporo-parietal junction that is regularly involved in explicit tasks is also activated during implicit belief processing. In line with this, Carruthers (2013) argues that even though people sometimes engage in fast decisions about mental states and other times deliberate on them extensively, this

[12] Both in ontogenetic and phylogenetic development it is an open question how ToM develops. For instance, contrary to what Apperly and Butterfill suggest, there is currently no evidence that anything like System 1 that can represent false beliefs would be shared with non-human animals (Call & Tomasello, 2008; Martin & Santos, 2016), unlike the system that represents exact numbers (Feigenson et al., 2002; Hauser, Carey, & Hauser, 2000).

simply reflects two different kinds of use of the same conceptual apparatus, similarly to other areas of reasoning (Carruthers, 2013). In addition, Carruthers questions the 'raison d'etre' of two systems, by pointing out that while positing two separate systems might enable 'simpler' mindreading occasionally, but this comes at a cost of an extra complexity of the overall theory of the domain (Carruthers, 2015b).

The separation of the two systems is facing another line of counterevidence from studies demonstrating spontaneous Level-2 perspective taking. Elekes and colleagues (Elekes et al., 2016) recently showed that adults spontaneously adopt a social partner's Level-2 visual perspective, if they know that she was paying attention to the aspectual properties of the object, because she was performing a task for which this information was needed. They found no such effect if the others' task was unrelated to aspectuality. According to their explanation, for Level-1 perspective taking it is enough to know that the other person has perceptual access to something. However, for Level-2 perspective taking it might be necessary to have some cue about the other person paying attention to the particular aspect of the stimuli that the perspective computation refers to. While this interpretation is appealing, it somewhat contradicts findings from another study, where in a similar paradigm they found spontaneous Level-2 perspective taking regardless of the partner's task (Surtees, Apperly, et al., 2016). However, in the latter case participants were asked to make eye contact before each trial, which might have facilitated perspective taking through increasing the joint nature of the task, even if in fact participants were performing it independently. While the details might be subject of different interpretations, these results suggest that under some circumstances Level-2 perspective taking might occur spontaneously.

In addition to the above considerations, there seems to be some circularity in the arguments behind what situations the two systems would be involved in. Since verbal tasks require explicit verbal reports, it is no surprise that it will require the involvement of a system that can have verbal access to belief representations. On the other hand, tasks that require fast reactions are said to be supported by a system that is capable of fast and efficient on-line computations. But since the type of access to the representation is measured by the type of answer required, the description becomes circular. What would determine which system is turned on, in the beginning of the observation of a scenario? In Rose Scott and colleague's work (Scott, He, Baillargeon, & Cummins, 2012) where they measured preferential looking to the pictures that match a story that is being told; which system is responsible for these reactions? Relatedly, in Clements and Perner's task (Clements & Perner, 1994), children were told a verbal story, and then their anticipatory looks were measured. Since in Clemens and Perner children did not pass the explicit task, which should be due to the fact that they do not have a System 2 yet; then System 2 cannot be responsible for their looking patterns. However, System 1 cannot operate on verbal content, as those are inherently propositional. Therefore such results cannot be easily explained by Apperly and Butterfill's 2-system view.

Finally, with regard to the psychological constructs the theory describes, Jacob (2013) pointed out that the concept of belief-like states is not clearly separated from the concept of beliefs. For instance, it is not clear how encountering differs from perception, as all the rules that are described to result in encountering seem to be the same that guide decisions regarding whether someone has perceived something. Jacob (2013) also criticized Apperly and Butterfill's (2009) claim that registrations can be false or incorrect. Jacob claims that for this, registrations have to be evaluated;

and iff at the time of evaluation they correspond to the state of affairs, then they are evaluated to be false (just like it is in the case of beliefs). However, Jacob argues that according to the theory, if the state of affairs does not correspond to a registration, it should not refute the registration. Therefore, he concludes, registrations are extensional and cannot be false. While Jacob is right on this point, this does not mean registrations could not serve as successful guidance for action prediction or interpretation. For an observer does not have to represent an agent's registration *as* being false or incorrect in order to use it for predicting the agent's actions. In fact, the same is true for beliefs that are represented as propositional attitudes. Even though propositional attitudes can have a truth value, this does not mean that interpreting a person's action based on a propositional attitude, one has to represent it *as* being false; as one may simply ascribe to this person a representation that happens to capture an incorrect state of affairs. Even the much-debated false belief tasks do not require such reasoning; they simply require predicting the agent's action according to her belief, or to report on her belief per se; none of which involves reflecting on this belief, or evaluating its truth value. One could argue however, that representing an attributed belief content that differs from one's own reality representation implicitly carries the information that the other's belief is false. But this poses problems for perspective taking, as if we accept that in perspective taking one represents the visual perspective and the resulting representation of the other person, in such cases the other's belief is different form one's own (and will need to be used to predict her actions), but nevertheless is not false.

Based on these considerations, the minimalist account faces some challenges, both with regard to theoretical arguments, and empirical data. While these concerns need to be addressed by proponents of this theory, the challenge to others is to

empirically test the claims made by this theory, and to provide an alternative of how ToM might operate. The minimalist account aims to give an alternative to the so-called nativist accounts of ToM, by positing an early developing, simpler system that only approximates full-blown ToM, and continues to operate in adulthood. As such, it is argued to have limitations with regard to the scope of mental state contents it can handle. Some of these limitations have been argued against based on theoretical grounds. For instance, Carruthers (2015) argues that positing such a system might seem cost-effective when describing the operation of this particular system, but it might mean extra complexity of the overall functioning of ToM (such as handling possible interactions between the systems). Other claims have been disputed based on recent empirical findings. For instance, since System 1 should not be able to handle Level-2 perspective taking, therefore Level-2 perspective taking should not occur spontaneously. But as discussed above, recent studies with adult participants found evidence to the contrary; suggesting that such allegedly more complex computations might happen spontaneously, and opening the possibility that even young infants might be able to perform such computations. Therefore more empirical work is needed to test the predictions of this theory in particular, and to characterize the nature of early mindreading in general.

Summary: possible limitations of early ToM abilities

Numerous accounts have emerged that argue for the lack of fully developed early mindreading abilities. Some claim that infants are incapable of any kind of belief understanding (De Bruin & Newen, 2012; Heyes, 2014a). We have argued against these on several grounds, and believe that to date no account has consistently explained the wide range of situations when infants were found to act according to an

understanding of others' mental states. But there is another reason to be skeptical that infants are completely 'mindblind', and don't understand anything of other's mental states. Namely, that infants (as well as adults) use Theory of Mind highly frequently in communication, for instance to understand pragmatics, or to establish a common ground (Rubio-Fernández, 2016).

Another group of proposals argues that infants possess some cognitive machinery that enables them to often approximate ToM abilities close enough to manage in a limited range of scenarios. Both Perner's mental file theory (Perner, 2016) and Apperly and Butterfill's minimal-ToM 2-system account (Apperly & Butterfill, 2009; Butterfill & Apperly, 2013) aim to provide an alternative to ToM as propositional attitude representations. We argued that while these theories seem to provide a viable description of early mindreading capacities; they face several theoretical challenges, and currently lack sufficient empirical support. Therefore assessing the predictions that follow from these is needed to evaluate such alternatives.

However, there are theories that have suggested that infants have some form of ToM abilities that constitute the core of later abilities, but nevertheless later get enriched (Carruthers, 2013; Leslie, 2000b; Westra, 2016). One candidate that might aid this process was discourse (i.e. conversations about mental states) that would guide infants in a gradual enrichment of (both implicit and explicit) mental state reasoning. Such proposals are consistent with the general finding that ToM development is influenced by mental state use of the mother, or growing up with an older sibling (Perner, Ruffman, & Leekam, 1994; Ruffman, Slade, & Crowe, 2002). While this is an intriguing possibility, it speaks to the question how ToM abilities

change over time through interaction with other domains and external factors, and does not address the cognitive mechanisms that enable representing mental states of others.

In sum, should one suppose the early emergence of ToM abilities, they will face two kinds of challenges. First, predictions from alternative accounts would need to be tested. Second, there is still a great need to fill in the gaps with regard to what cognitive mechanisms enable representing others' mental states. In the following sections we will entertain the hypothesis that infants are capable of representing others' mental states in a form that can be the basis of a fully functional ToM system, which we will call a one-system account. We first provide arguments as to why one should entertain this possibility, and then continue with some predictions and open questions that may lead us to a better understanding of early mindreading abilities, together with describing the challenges such an account faces in light of the alternative theories.

1.2. Mechanisms of early ToM abilities

Proposing a one-system ToM with multiple components

Several theorists have argued that core mindreading abilities are present from birth, and continue to operate into adulthood. Most of these theories have challenged the necessity to posit different concepts to describe infants' ToM abilities. The conceptual change accounts, arguing for discontinuity, seem to differ in the reasons behind not granting children a full-blown ToM. Perner claims that the ability to form

the necessary links is missing; which might or might not turn out to be true empirically, but in principle it can be argued that linking is a distinct cognitive mechanism that establishes sameness of reference, whether between two regular files or one's regular files and vicarious files. However, Apperly and Butterfill argue backwards from the claim that representing propositional attitudes are cognitively demanding because they *could* be used in complex ways. First, this is more a claim about the rapidness of computations (and the basis for the assumption that for instance Level-2 perspective taking should not happen automatically), rather than a claim about infant cognitive abilities. Second, as Carruthers points out, just because propositional attitudes (or metarepresentations of others' belief contents) have the *potential* to exhibit these features, it doesn't mean they do it under any circumstances. Indeed, there is nothing inherently cognitively demanding about entertaining a representation and use it in some possibly simpler way, just because the same representation could be used in much more complex cognitive processes. Very blatantly put: if I use in a very simple sentence some English words that I know the meaning of; I can probably use them in their right meaning, even though they might be used in Edgar Allan Poe's poems that undoubtedly can be more sophisticated then my sentences. However, this does not mean that I don't have an understanding of these words, or would not be able to use them in a fast and efficient manner just because they *could* be used in a much more elaborate way.

It is certain that ToM abilities go through a great amount of development from infancy, through childhood, to adult life. Early nativist-modularist theories have been accused of being anti-developmental, but as Leslie (2000) points out, assuming an innate or modular basis of ToM does not exclude the role of learning. We do not wish to make a point here about the possibly modular nature of ToM. The aim is to

emphasize that in fact all theories are nativist to some extent, as they must agree that infants are born with *some* abilities that at a certain point in development then enable them to represent mental states. What is at stake is whether the concepts that ToM operates with are the same from early on, or they go through some radical change at some point in development. Carey (e.g. Carey & Johnson, 2000) distinguishes between knowledge enrichment and conceptual change. It is suggested that a combination of metarepresentational abilities and general computational power together lead to the construction of metacognitive knowledge that in turn enables conceptual change. Taking the case of children with Williams syndrome, Carey and Johnson (2000) point out that these children successfully pass false belief tasks with some delay around the age 6, but never make some conceptual changes that typically developing children make a few years later (such as in the domain of biological knowledge). Based on these data they argue that ToM is unlikely to require conceptual change, as children with Williams syndrome acquire ToM, although they seemingly have very limited cognitive abilities that impede them to perform conceptual change in general.

To date there is no good evidence that infants in principle should not be able to form propositional attitude representations. And as Leslie (2000) argues, it is unclear why there should be an innate capacity that then needs radical revision, when there could be an innate capacity that simply needs refining. Leslie originally raised this argument against early theory-theorists, who argue that ToM *is* actually a theory that the child possesses (Gopnik & Wellman, 1992; Wellman, 1992). Such theories claim that there is an innate theory the child originally is endowed with, but that is limited in its capacities, as it initially states that "beliefs are copies of reality", but then by the age of 4 it is changed to "beliefs are representations of reality" (Gopnik & Wellman,

1992). Yet, there is no prima facie reason why the child could not possess the correct theory in the first place. Analogously, there is no prima facie reason why the child could not represent mental states as propositional attitudes.

In fact, representing propositional attitudes might be simpler than entertaining some other form of representation. In the case of propositional attitudes, the content of this attitude is a representation that the infant herself can entertain that needs to be embedded into a structure that specifies the corresponding agent and attitude. However, to entertain registrations the infant would need to compute a representation that differs in its form from the infant's own representations (Carruthers, 2013).

Nevertheless, whether an infant is capable of representing others' beliefs in the form of propositional attitudes may be an unnecessarily strict criterion of a fully operational ToM. One might argue that it is possible to form metarepresentations of others' mental contents without representing these contents necessarily in a propositional format. Yet the metarepresentations infants can form about others' belief may carry the crucial characteristics that are credited to propositional attitudes. Specifically, it may apply to them that the content cannot be substituted arbitrarily with other representations of the same referent. As such, infants could in fact form representations that are in line with understanding aspectuality of mental states, which infants have been argued not to be capable of (Butterfill & Apperly, 2013; Perner et al., 2011). Therefore we take the critical question to be whether infants can form metarepresentations of others' mental states that fulfill the above criteria.

Such mental state representations could have a similar syntax as propositional attitudes: they may consist of an agent, an attitude, and a content; where these elements could be changed flexibly. Indeed, there is some suggestive evidence that parts of this structure can be represented as placeholders. For instance, the agent whose mental state is computed is not always represented, or rather, the mental state might not be bound to the agent (Kampis, Somogyi, Itakura, & Király, 2013). Additionally, infants could open placeholders for the content of beliefs and only later fill in the specific belief content (as proposed by Kovács, 2015). Finally, while to our knowledge this has not yet been directly tested, infants can mostly likely represent various attitudes *of* the same agent *to* the same content, such as 'desire' or 'believe'; which would mean that the attitude element can also be changed flexibly. Most studies design scenarios where the prediction of an agent's action relies on a preference for, *and* a belief about, an object; however this itself does not show that the two representational structures could be flexibly turned into each other by changing only the attitude element.

This characterization of the format of attributed beliefs is captured by the notion of 'belief file' (Kovács, 2015). Belief files describe a representational format that contains the critical representational elements enabling online belief tracking. It contains two variables: an agent and a content. Both variables can be modified independently from each other, which enables fast updating and modification of the elements. As such, this format enables rapid computation on any belief content – rather, the limit of the speed and effort of computing a particular mental state content would depend on the difficulty to calculate and represent such content itself. This fits with Carruthers' claim that Level-2 perspective taking is difficult in part because it requires additional processes such as mental rotation (Carruthers, 2015a). This

characterization of belief representations can give a possible solution to how mindreading can be both sophisticated and fast at the same time. When belief computation is triggered (e.g. through the detection of an agent, who is a potential belief-holder), the entire skeleton of belief file is set up. From then on, individual parts can be more or less specified, for example the content can be under-specified in cases when the observer does not know the specific content of the other's belief (e.g. Sally knows that Anne thinks *something* is in the box). If some element needs to be modified, e.g. the content needs to be updated; the individual elements can be accessed and modified without having to start over the process of belief attribution.

The notion of belief-files fits well with accounts that describe ToM as a rather multi-componential process (Carruthers, 2013; Kampis et al., in press; Kovács, in prep.). The common suggestion is that belief computations are a result of the orchestrated functioning of many different cognitive systems. Computing the elements of belief representations already require several different processes to take place (such as detect agents as potential belief holders, and compute belief contents); which in turn may interact with infants' own reasoning and planning systems, in order to then predict or explain the behavior of another agent (Carruthers, 2013). This idea well explains much of the development that might happen from infancy to childhood. Through the enrichment of other systems, mindreading becomes more and more sophisticated, and among others, becomes part of discourse (Westra, 2016). In line with this, according to Leslie's account it is the capacity to represent propositional attitudes that is at the core of an innate module that humans possess (Scholl & Leslie, 1999). Leslie claims that there is a certain part of ToM that operates in a modular way, but stresses that this is consistent with the possibility that these modular processes work together with other modular, or more domain-general

processes. Moreover, a modular notion of this sort is also compatible with development *within* a system; as for example there might be development with regard to the kinds of mental states one can entertain.

It seems therefore theoretically plausible, that mindreading abilities share their main characteristics throughout the lifespan. According to this hypothesis, humans possess an innate capacity to represent metarepresentations of others; mental contents, which constitutes the core basis of Theory of Mind. As we pointed out earlier, such an account positing an early developing ToM system will face two kinds of challenges. First, it needs to give a story of how ToM processes might operate, and see whether the predictions hold against empirical evidence. Second, it will need to address predictions from theories suggesting specific limitations of mindreading. In the next part we will turn to outlining some of the mechanisms that might underlie representing others' beliefs; and then discuss some of the predictions that follow from this possibility in comparison to alternative accounts.

1.2.1. Meta-representing belief contents in infants

Much of the debate on early belief reasoning abilities concerns the question of whether infants are able to entertain metarepresentations with the content of attributed representations. Whether infants can be granted such capacities partially depends on what definition one relies on. Perner (e.g. Wimmer & Perner, 1983) refers to Pylyshyn's (1978) description, and states that according to Pylyshyn if someone has the ability to impute mental states to oneself and to others they "not only have a representation about a state of affairs (x) and stands in certain relationships to these representations… but also represents these relationships *explicitly*" (Wimmer & Perner, 1983, p. 104, *italics added*). This appears to entail some sort of *metacognitive* reflection, which is also suggested in Perner's notion of linking regular mental files to each other and to vicarious files (Perner et al., 2015) where conscious reflection on the link is required for metarepresentations. However, this need not be the case in order for linking to work, and it is not necessary for metarepresentations to fulfill their role. What is necessary is for the cognizer to use the metarepresentations in their appropriate role. For instance, one uses their own representations to *plan* their *own action*, and should use metarepresentations of an agent's mental content to *predict* the corresponding *person's actions*. Indeed, if one agrees to exclude alternatives that do not adhere to any sort of mindreading but some lower-level mechanisms; what remains is the option that infants indeed use these attributed beliefs *as* they ought to be used. In fact, Pylyshyn did not claim that the relations needed to be represented explicitly, but merely that they needed to be represented in the first place: "The studies P&W report are an attempt to show that not only can the chimpanzee be in the kind of relation to a representation that Russell called "relations

of propositional attitude" (e.g., believing that B, wanting it to be the case that B, expecting that B, wondering whether it is the case that B, or even considering what would happen if B were the case), but also that he can *represent these relations themselves* - or specifically that he can represent that *other organisms are in these relations to their representations*. In other words, the studies argue that we must not only posit beliefs in order to account for certain of the chimpanzee behaviors, but also beliefs about beliefs and intentions of others (and, presumably, about the chimpanzee himself)." (Pylyshyn, 1978, p. 593, *italics added*). From this it does not follow that in representing propositional attitudes (or in metarepresentations of other' mental contents in general) the relation needed to be explicit, simply that the relation itself needs to be represented. But this criterion is fulfilled by representing the agent-attitude-content relationship, even without explicit access to the link (or any other part).

Based on these considerations, the present thesis will assume that representing mental states of others entails metarepresenting the representations one attributed to another agent, and using these according to the general function of representations, such as the fact that they influence one's behavior. What cognitive mechanisms may enable forming such metarepresentations? As Carruthers (2013) pointed out, when infants represent others' propositional attitudes, the content of this attitude is a representation that the infant entertains as a primary representation, which then needs to be embedded into an appropriate agent-attitude-content structure. In line with this, Leslie (1987) in his work describing a cognitive model of pretense (make-believe play observed in toddlers, such as pretending that a banana is a telephone) argues that in pretense the primary representation of an object is copied into a 'metarepresentational context.' Relatedly, Sperber (2000)

proposed that the most cost-effective way for a cognitive system to handle metarepresentations would be if any representation could also serve as the content of a metarepresentation. All three of these proposals involve some form of 're-use' of a primary representation when computing others' mental states.

The present thesis proposes that infants can exploit their representational machinery that is responsible to encode the world from their own perspective, to metarepresent others' representations about the same referents; this would provide a powerful tool to potentially attribute any mental state to others that the infants themselves can entertain. Such a possibility in turn would predict that infants should succeed on a variety of false belief tasks, among others ones involving the attribution of abstract representations (e.g. quantifiers), aspectuality, or absent referents. Moreover, this would enable to make predictions regarding the attributes of the cognitive processes involved. For any cognitive system that is involved in computing primary representations about a certain referent, should also be involved in metarepresenting the same as the content of a mental state representation.

Indicators of shared mechanisms for own and attributed representations

The idea that one's own cognitive systems are involved also when thinking about others has appeared before. A similar notion was behind the simulation theories of mindreading that argued that we come to understand others' mental states through simulating the events internally, reasoning from a pretend scenario, and possibly involving the 'mirror neuron' system (Gallese & Goldman, 1998). The mirror neuron system is a set of neurons that was initially discovered in the premotor area of the

macaque monkey, to respond both when a specific action was executed by the monkey, and when the monkey observed someone else to perform this action; and soon after similar results were found in humans as well (for a summary on the mirror neuron system in monkeys and humans see Rizzolatti, Fogassi, & Gallese, 2001). The role of this system was suggested to 'retrodict' the other person's mental state, moving backwards from the observed action that is mirrored in one's own motor system (through direct matching). However, this view was later heavily attacked from both theoretical and empirical angles. First, neuroimaging evidence suggests that contrary to what simulation theories should predict, the brain areas involved in mindreading are not sensitive to the similarity between oneself and the observed person (Saxe & Wexler, 2005). On a different line of arguments, others have shown that infants attribute beliefs to geometrical shapes (Surian & Geraci, 2012) and cartoon characters (Kovács et al., 2010); they can compute goals of biologically impossible actions (Southgate et al., 2008), and show motor activation during prediction of nonexecutable actions (Southgate & Begus, 2013) in which cases direct matching of the perceived action would not be possible, as the actions are not part of the observer's own motor repertoire. These considerations made it unlikely that such simulative processes underlie the understanding of actions. Indeed, Csibra argued that the "primary function of action mirroring is not action understanding in terms of goals but *predictive* action monitoring… action understanding may precede, rather than follow from, action mirroring" (Csibra, 2007a, p. 436, *italics added*). In accordance with this, others have found that motor activation in fact *precedes* the observation of a movement in human adults (Kilner, Vargas, Duval, Blakemore, & Sirigu, 2004) as well as in infants (Southgate, Johnson, Osborne, & Csibra, 2009).

These results support the notion that one's own representational machinery can be recruited in social interactions. However, unlike suggested by the simulation theories, it is not the recruitment of these mechanisms that *enables* the attribution of mental states, but vice versa: once a mental state is computed (based on an observed action), the need to *generate* a mental state content, and the possibility to generate action prediction based on the mental state, *emerges*. This latter process is then aided by one's own cognitive systems. One example that supports this is the involvement of the motor system in action prediction. Indeed, electrophysiological findings suggest that infants recruit their motor system during the prediction of another person's actions, but only when there is a possible goal that infants can infer to be the intended outcome, and hence the action can be interpreted as goal-directed (Southgate, Johnson, El Karoui, & Csibra, 2010). More recently, Southgate and colleagues (Southgate & Vernetti, 2014) have found that infants show motor activation not only when they perform an action, but also when they predict someone else's action based on her false belief. The authors measured sensorimotor alpha suppression as an indicator of action prediction, and found that infants show such activation when an agent falsely believed that a ball was present in a box (and therefore was likely to reach for it), but did not show the activation when she falsely believed the ball *not* to be in the box (and therefore probably would not reach for it); regardless of infants' knowledge of the true state of affairs.

Others investigated what cognitive processes enable the tracking of how other people around us interpret communicative acts (Rueschemeyer, Gardner, & Stoner, 2015). In an electrophysiological study they presented participants with sentences, in the presence of a confederate. The sentences were either semantically plausible, or implausible for *both* the participant and the confederate; or plausible for the

participant but implausible for the *confederate*. They measured changes in a specific event-related potential (ERP) component, the N400, which was previously found to accompany the processing the difficulty of semantic integration. They confirmed a larger N400 component when participants themselves heard the semantically implausible sentences; and crucially, they found a similar effect when the sentences were plausible for the participants themselves (as they received disambiguating information), but implausible for the other person – which they termed the 'social N400 effect'. The authors stress that in this paradigm calculating the other's perspective was not spontaneous, as it was included in participants' task to judge whether they understood the sentences and whether the confederate understood them. Nevertheless, these results are compatible with the possibility that participants represented the sentences and their semantic content from the other person's point of view; and used their cognitive mechanisms that are involved in detecting semantic mismatches from their own perspective, to detect such semantic mismatches attributed to the confederate.

In most belief tracking scenarios, and certainly in everyday life, tracking others' mental states is intertwined with observing actions involving objects (whether social, or physical objects) in the environment. Hence, representing the content of a mental state, and to use this to interpret or predict someone's action, likely involves the recruitment of various cognitive processes that interpret and track the world around us. It is therefore likely that the operation of multiple systems can be observed in association with belief representations. The above examples were the first few that investigated such processes; but it is that many other similar effects should occur. With regard to infants, we now we have ample evidence that from early on they can reason about a wide range of phenomena, involving both physical (e.g. physical

causation) and social (e.g. helping/hindering) events. If encoding particular events from infants' own and from someone else's perspective recruits analogous cognitive processes, then we should observe similar characteristics for the two. The social N400 effect, and the sensorimotor alpha suppression in action prediction are two such examples; and the present thesis hopes to contribute to the understanding of how infants' cognitive systems take part in mental state attribution, by investigating signatures of ascribed object representations (in Chapter 2).

In addition to making predictions about the features of the mechanisms involved; building on shared representational resources for entertaining one's own representations and someone else's is also consistent with the growing body of evidence that shows that people's behavior can be modulated by someone else's visual perspective, or beliefs about a situation.

The modulation effect: How belief contents influence one's own behavior

The modulation effect originates broadly from two studies that in parallel aimed to investigate the ease with which human adults compute someone else's perspective. Samson and colleagues (Samson et al., 2010) were interested in whether people occasionally take someone else's visual perspective spontaneously. They presented participants with images where a human avatar was standing in a room, facing one of the walls, and on the walls there were various number of discs. Participants' tasks was to judge the number of discs from (i) their own, or (ii) the avatar's perspective, while the two perspectives were congruent (they could see the same number of discs) or incongruent (the avatar did not see some of the disks as they were behind his/her back) with each other. Results showed an egocentric intrusion, whereby participants'

own perspective slowed down judgments of the avatar's perspective in incongruent trials. Interestingly, there was also an altercentric intrusion, which was apparent in slower judgment of one's own perspective on trials where the avatar's perspective was incongruent with theirs. This effect was still present (albeit smaller) in a different group of participants who only had to judge their own perspective, without receiving any instruction to compute the avatar's perspective. The authors took these results as evidence that participants spontaneously and rapidly computed the other person's perspective, which interfered with their own perspective judgments. However, they also stressed that "for simple perspective-taking problems, the content of someone else's perspective does not require the complex representational apparatus often discussed in the literature on theory of mind [...] In an ecologically useful range of circumstances, adults may be influenced by much more low-level computation" (Samson et al., 2010, p. 1264). This latter conclusion is far from obvious, however. First, it is unclear what 'high' and 'low' level computations mean in this case. But more importantly, the fact that the avatar's perspective modulated participants' own judgments suggests that participants computed the content of her perspective. Moreover, this attributed representation has to be of a similar format than participants' own representation; as suggested by the fact that both could feed into participants' action planning. Therefore what is of importance here is not whether this is 'low-level' or 'high-level' computation; but that such an interference effect is only possible if the two representations share some fundamental characteristics.

This latter hypothesis motivated the study by Kovács and colleagues (Kovács et al., 2010). In an object-detection task they investigated the effect of another agent's belief on someone's own decisions. Participants watched short animation sequences,

where an agent (an animated smurf) observed a ball first going behind an occluder. Then the ball either rolled out from behind the occluder and exited the scene ('P-' condition, as the participants believed the ball to be absent), or left but then rolled back behind the occluder ('P+' condition); and as they varied when the agent exited the scene, he either ended up believing the ball to be present (A+) or to be absent (P-). This resulted in altogether 4 conditions, two of which were 'false belief' conditions, as in half of the cases the agent's belief did not match the participants' belief. In the end of this sequence, the occluder always fell, and the ball was either present or absent (50% of the time it was the opposite outcome to the participants' belief). The task for adult participants was to press a button if the ball was behind the occluder, and not to press if they didn't see a ball when the occluder fell. As predicted, participants were faster to press the button when both the agent and they themselves expected the ball to be behind the occluder (P+A+), compared to when neither expected it to be there (P-A-). However, in half of the cases when they did *not* believe that the ball should be there, the agent nevertheless had a false belief that the ball is present (P-A+). In these cases, participants were again faster to press the button, suggesting that the agent's belief that the ball is present facilitated the participants' button press. They concluded that the computed belief representations seem to be similarly accessible to subsequent cognitive processes, as participants' own representations. Analogous results were obtained with infants, where in a looking-time version of the task infants were surprised (as suggested by longer looking times) when the ball was absent after the occluder fell, when they and the agent thought it was present (P+A+), but were not surprised about such outcomes when *neither* they nor the agent believed it to be present (P-A-). However, they again looked longer at outcomes where the ball wasn't there, when *they* knew the ball should be absent, but

the *agent* believed it to be present (P-A+); again suggesting that infants' looking times were modulated by the agent's belief that the ball is behind the occluder.

Similar findings were obtained with a continuous measure in adults, where the task was to decide where the target object is hidden between two locations, and to move a computer mouse to this location from a set starting point (van der Wel, Sebanz, & Knoblich, 2014). Results showed that the movement trajectories of the mouse were influenced by another agent's belief about the location of the target, suggesting that participants incorporated the ascribed belief in their online decision process. Several other studies found such modulation effects, among others with children (Surtees & Apperly, 2012), and even in live Level-2 perspective taking scenarios, as discussed earlier (Elekes et al., 2016; Surtees, Apperly, et al., 2016; but for absence of Level-2 perspective taking with computerized stimuli see Surtees, Samson, et al., 2016).

These results are well explained by the possibility that belief contents are handled with the involvement of the same cognitive systems that represent the environment from one's own perspective. This social modulation effect could be on one hand considered an 'error' within the ToM system, where the ascribed representations 'intrude' into one's own planning processes. On the other hand, as Kovács et al. (2010) point out, such rapid and easy availability of access to others' beliefs might have the benefit of enabling efficient social interactions as well as social learning, the latter being especially important in infancy. However, the current evidence is limited to a narrow range of behaviors that requires relatively minimal action planning, in fairly simple scenarios. If this effect is indeed robust, then we should observe it in a variety of situations, including ones that may require more

'active' behavior, such as going to a target location, or searching for a hidden object. Moreover, to date only one study (Kovács et al., 2010) investigated this effect with infants, therefore one goal of this thesis is to widen the scope of situations in which the modulation effect occurs in infancy.

An obvious advantage of being able to utilize various cognitive mechanisms to calculate others' mental state contents is that in principle it enables ascribing any mental state content to others that individuals can themselves entertain. The last decades of research have characterized infants' understanding of various social and physical phenomena around them. In contrast, research on mindreading has investigated a fairly narrow range of phenomena for which infants can ascribe mental states. Building on our knowledge of infants' abilities to represent the world around them enables us to make predictions on the types of mental state contents infants should be able to handle, in order to better characterize early mindreading abilities. One domain that is well researched in infants is their knowledge in the realm of objects. In the next section we describe some characteristics of object cognition in infants, with an eye on their potential use in belief tracking.

Object representation capacities in service of belief representations

In order to successfully navigate their surrounding and to learn about the world around them, infants need to master a great amount of knowledge about physical objects and events. What happens if they lift or drop an object? Or if it rolls away? Or if it goes behind something? When is it justified to look for an object somewhere? Do objects cease to exist when they are not visible anymore? Infants encounter such

dilemmas on a daily basis. The system of object representation has been suggested to be one of the core systems that are part of humans' genetic endowment (e.g. Spelke, 1994, 2000).

Object cognition can be grasped from various angles, such as describing objects with regards to the events that they are part of (Baillargeon, 2004), or their conceptual representation (Spelke, 1990). Leslie and colleagues (Leslie, Xu, Tremoulet, & Scholl, 1998) explained object cognition in relation to visual attention, to provide an account of the cognitive system that may enable the tracking of objects in the environment from early on. Object-based visual attention was suggested to indicate a basic mechanism that provides units for characterizing the limitations of processing visual information (Pylyshyn, 2001). The visual system assigns indexes to objects, and these indexes enable tracking the objects continuously through space and time, even during brief occlusion of the object (Scholl & Pylyshyn, 1999). An index itself does not store any feature information of the object, but it enables the opening of an object-file in short-term memory that can potentially store such information. Object-files are linked to the objects of the environment through the indexes. In line with this system being operational from early on, infants as young as 3-5 months of age were shown to have expectations that objects will continue to exist after occlusion (Baillargeon, 1986; Baillargeon & DeVos, 1991; Baillargeon, Spelke, & Wasserman, 1985). Others explored the neural bases of maintaining such object representations, and found that in six-month-old infants a specific brain activation, namely gamma-band oscillatory activity in temporal regions, accompanied events when previously observed objects were occluded from the infants (Kaufman, Csibra, & Johnson, 2003). Moreover, they observed similar activation when infants faced a mismatch between the presented events and their maintained representation, e.g. at

unexpected disappearance events when an object should have been at a revealed location, but in fact it was absent. They also obtained congruent looking-time data, where infants looked longer at unexpected disappearance events than to expected disappearance events. They argued that this pattern is consistent with the fact that infants detected the mismatch between their sustained representation and the visual input, and they were trying to resolve this incongruence, which resulted in the longer looking times. In a subsequent study they found that such gamma-band activation adheres to some suggested principles of the object-index system. Specifically, they found that there was an increase in activity only in case the object disappeared in a manner that was consistent with its continuing existence (Kaufman, Csibra, & Johnson, 2005). These results are consistent with previous findings from studies with human adults showing continued tracking of an object when it disappeared because it was being occluded, but not when the object disintegrated; suggesting that in the latter case the object-index was discarded (Scholl & Pylyshyn, 1999).

Such abilities to track objects continuously even in the case of occlusion, are crucial for infants to learn about their environment. However, they are also of importance in social interactions. How would one know what the other person points to? Or what they are looking for? Or where they will go to acquire an object? It is necessary to track what others have attended to, and therefore what they might know about objects' existence, location, and so on. Indeed, infants build on this understanding when they take others' perspective, and when they track their beliefs. For instance, every location-change false belief scenario involves the tracking of an object from being visible, to being hidden to various locations and then emerging again. The fact that infants have expectations about others' actions in such scenarios supposes that they assume the other's individuation, object tracking, and object

permanence operate the same way as theirs. For instance, in the pioneering study by Onishi and Baillargeon (2005) in one condition infants observed the agent with the object in her view, and then the object being placed in a box (therefore was not visible anymore). Following this, the object came out of the first box (and hence became visible again), and entered the other box (and hence again went out of view). During these events, the actor was present. Next, in her absence the object moved back to the first box. Successful representation of her belief of the object relied on this case on attributing to her (i) the individuation of an object (ii) tracking the object to location A, (iii) maintaining the object representation in location A, (iv) tracking the object go from location A to B (v) maintaining the object representation in location B until she plans her action to reach. Moreover, similarly to how the infants in Kaufman et al. (2003) showed both (a) increased gamma-band activation when they detected incongruence between their sustained object representation and the visual input, and (b) looked longer during the observation of such events, possibly due to the attempt to resolve the above incongruence; infants in Onishi and Baillargeon's (2005) study looked longer when the agent reached to the location that was incongruent with her representation of the object.

This raises the question, what mechanisms enabled infants to track the object from the actor's perspective during the observations of the above steps (i)-(v). A representational understanding of beliefs would entail that infants in such scenarios attribute the representation of an object to the other person. Earlier we proposed that for representing other's mental contents infants recruit the same representational machinery that was found to support representing the environment from their own perspective. In Chapter 2 of this thesis we will investigate, among other questions, a prediction that follows from this proposal: the possible

involvement of gamma-band oscillatory activity (cf. Kaufman et al., 2003, 2005) in object-related perspective taking and belief tracking scenarios.

In most cases there are multiple object around us, and often it is relevant to track more than one of them. Infants not only individuate and track one object through occlusion, but they also use spatiotemporal information to establish numerical identity, for example to infer that spatiotemporal discontinuity means two numerically distinct objects (Carey & Xu, 2001). In a looking-time study 5-month-old infants after observing two objects of the same appearance go in sequentially behind an occluder, were surprised when the occluder fell and revealed only one object (Wynn, 1992). As such, infants can perform simple arithmetic operations involving a small number of items: through the assignment of distinct indexes to each object this system implicitly represents information about the *number* of objects that are tracked at any given time. This tracking is possible up to a limited number of 3 or 4 items (Scholl, Pylyshyn, & Feldman, 2001). Numerical equivalence between two arrays can be established through one-to-one correspondence of the indexes. As infants' computations were shown in such scenarios to also be limited to the set-size signature of around 3 items, the object tracking system was suggested to underlie the quantifications involving small object arrays (Feigenson & Carey, 2005; but see Cordes & Brannon, 2008).

To assess whether infants can use their object-file system to determine numerical equivalence between a set of object-files they store in memory and a set of objects in the world, Feigenson and Carey (2003) used a manual search paradigm (adapted from Van de Walle, Carey, & Prevor, 2000), and presented 14-month-old infants with an opaque box, followed by a certain amount (2, 3 or 4) of objects placed inside. Next,

infants were allowed to retrieve these objects. Measuring the duration of infants' manual search in the box showed that infants keep searching for an object if they have not yet retrieved all the objects that they previously saw being placed inside. Specifically, infants searched longer in the box if one of the objects was still inside, compared to when all objects that were placed in the box were already retrieved. This was the case when infants counted back from 2 or 3 objects, down to 1 object remaining or all objects retrieved. However, infants' tracking of the objects broke down at 4 items; when counting back from 4 they did not search longer when an object still remained in the box, compared to when all 4 were retrieved. Based on this specific pattern of success and failures (tracking up to the set-size limit of 3 items.) Feigenson & Carey (2003) concluded that infants in this case represented the objects with the help of the object-tracking system that was previously found to show such limitations.

Infants possess powerful cognitive mechanisms that enable representing and tracking objects around them in relatively complex scenarios. In contrast, our knowledge of how much they can reason about when it comes to others' mental states involving objects is fairly limited. Infants can individuate and track multiple objects at a time, and through their first year of life they develop the ability to store more and more information about these objects (Leslie et al., 1998; Wilcox, 1999). Around their first birthday, when their motor abilities enable them, they can use these abilities to actively make choices and explore the environment (Feigenson & Carey, 2003; Feigenson et al., 2002). If as suggested in the present thesis, infants can exploit these resources in the service of mindreading when representing others' belief contents, then the range of scenarios infants can handle may be much wider than what has been investigated so far. We propose that any particular infant should

be able ascribe to others representations that she herself can entertain. For instance, unlike the majority of scenarios that investigate false belief understanding that involve one single object, infants should be able to handle false beliefs that require tracking multiple objects, possibly up to the limits of their object tracking system. Moreover, if an infant can handle specific information necessary to infer and represent object identity, they should be able to attribute to others such representations to another agent as well.

In other words, unlike the minimal-ToM account (Butterfill & Apperly, 2013), this account would predict that in fact infants should *not* show *arbitrary* limits with regard to the types of belief contents they can attribute to others. Rather, such limits should be systematic with regard to the cognitive system they involve, and would come from the inability to compute certain representations or track particular kinds of information at certain ages in development, or under cognitive load; both in infants, and in adults. The present thesis aims to contribute to the characterization of early ToM abilities with an empirical investigation of the above outlined predictions on attributing representations involving multiple objects or object identity (in Chapters 3 & 4); and when applicable, contrasts it with the minimal-ToM account that posits limits with regard to belief contents, regardless of infants' own abilities.

In the present section we referred to the construct of object-files to describe some of infants' abilities to represent objects around them. The idea of a file-like structure as the organizational unit of attributed representations emerged in several recent accounts (Kovács, 2015; Perner, 2016; Recanati, 2012). These offer somewhat different characteristics of the structures of the files, but more importantly, they offer an explanation on different organizational levels. Object-files were introduced to offer

a mid-level representational 'layer' between perceptual representations of unbounded features (Kahneman, Treisman, & Gibbs, 1992), and later, more conceptual representations (e.g. object kinds). One object can have only one index attached to it (and hence one object-file). Perner's (Perner et al., 2015; Perner & Leahy, 2016) mental files, however, allow for multiple filing, where various files attached to the same object (referent) express different aspects (senses) of the same referent. Combining these, here we propose that mental files might not be linked to the objects themselves, but to the object-file, which is then indexed to the external object. This possibility would solve a problem that our suggestion would otherwise face: namely, that object-files are individuated by spatiotemporal indexes, therefore *one* object cannot have *two* indexes; which makes object-files unsuitable to be 'attributed' to others. However, there is no suggested limit on the number of mental files to be opened for an object; therefore it allows for 'vicarious files' to serve the role of attributed object-representations. A possible mechanism would be that if there is a need to ascribe a representation to someone else, then a mental file will be opened for the other person. This would make an intriguing prediction. Namely, the suggested limitation for the object-file system comes from the limited amount of indexes that can be assigned at a time. However, if the attributed representations are represented in the mental file system, then they should not have to adhere to the same limitations; allowing, in principle, to sustain a larger amount of vicarious files than object files. In other words: through this mechanism it would not be possible to represent one person's (not even one's own) representation of 5 objects, but it could be possible to represent 5 people's beliefs about one object[13].

[13] There could be additional constraints, such as working memory limitations that may restrict such attributions. We will return to this issue in Chapter 3.3.

Finally, the belief-file refers to the entire belief construct, where the content is one element, alongside the agent variable. Therefore mental files could serve as the content of the belief-file; and possibly also the agent can be represented through a mental file. As such, these different constructs serve to characterize the various elements necessary to ascribe beliefs to others. These different mechanisms together may enable the formation of mental state representations in infants. While such a detailed analysis is lengthy and complex; each element can in fact operate in a fast manner, thus potentially serve as a mechanism for automatic belief attribution.

Summary: cognitive underpinnings of representing belief contents in infants

In the present section we explored the possibility that ToM abilities in their core format are present from early on. First, we discussed arguments that support an early developing (possibly innate) capacity to represent mental states through metarepresenting others' representations. We briefly outlined a componential view of ToM where multiple cognitive processes compute the distinct elements of a flexible representational structure, the 'belief-file'. This structure has a placeholder for an agent and a belief content. Focusing on the latter, we then suggested a possible mechanism through which belief contents are handled; where one's cognitive mechanisms that represent the environment are also recruited to compute representations attributed to others. We argued that this possibility explains well the modulation effect, where ascribed representations influence one's own behavior. Additionally, we pointed out that the modulation effect needs to be further explored in more active scenarios. Finally, we described two predictions that follow from this suggested mechanism. First, signatures of representing primary representations

should also accompany representing them as belief contents. Second, infants should be able to handle belief representations with a variety of contents that they themselves can entertain (such as beliefs about multiple objects); and should not show arbitrary limitations with regard to the types of belief contents (such as beliefs about identity), but rather limitations that are defined by their own representational abilities. Together, these claims make strong predictions on how belief contents are represented in infants. One aim of the present thesis is to explore these predictions to gain a better understanding of early ToM abilities.

1.2.2. Automaticity and flexibility of ToM mechanisms

Humans compute others' perspectives even when they are not asked to do so, to the extent that others' mental states can influence one's own behavior (Kovács et al., 2010; Samson et al., 2010). Such findings raised the question of automaticity of mindreading, and inspired studies investigating the circumstances under which we compute others' perspective and mental states. Primarily, research has focused on whether people compute others' mental states without overt instruction (Apperly, Riggs, Simpson, Chiavarino, & Samson, 2006; Cohen & German, 2009; Kovács et al., 2010; Samson et al., 2010; Schneider, Bayliss, Becker, & Dux, 2012; van der Wel et al., 2014). The picture that seems to be emerging is that belief reasoning can happen automatically (independent of one's goals, whether they are implicit or explicit goals; adopting terminology from Carruthers, 2015a), or at least spontaneously (independently of external prompt or explicit goal; but possibly depending on implicit goals and therefore involve executive resources). Moreover, as we reviewed

in the introduction, there is ample evidence in infants with spontaneous measures that suggest instructions are not necessary for mindreading to take place.

There are, however, less explored aspects of the question of when mental state computations take place. First, another way to characterize spontaneous mindreading, is to analyze the time point *when* belief computations take place within a particular situation. Another open question is under *what circumstances* do belief computations happen with regard to the relation between one's own representations. Additionally, the flexibility of such processes is largely unexplored. Within a situation one often has to incorporate changes into their own representation. Similarly, one may have to update the representation attributed to the other person. Currently we lack an understanding of how these changes happen, under what circumstances do they happen, and *whether* they always take place.

Belief ascription in perspective taking

Recall the three philosophers' challenge (Bennett, 1978; Dennett, 1978; Harman, 1978) to Premack and Woodruff's (Premack & Woodruff, 1978) argument. The false belief task was developed because one cannot represent from their own perspective a state of affairs as reality that she does not hold to be true, therefore if the other's belief is mistaken, then one necessarily has to represent it as a distinct representation from one's own. Relatedly, many of the tasks aiming to assess spontaneous mindreading use false belief scenarios. However, there are many other cases when the other's epistemic access will be different from our own, and in order to interpret or predict her actions, we need to attribute a particular representation to her. In fact, perspective taking scenarios are such cases; and we argued earlier that the task by

Samson et al.'s (2010) suggests that participants computed the other's representation which then had an effect on their own behavior (as for instance, computing the line of sight between the other person and an object would hardly cause interference with one's own representation). Theoretical considerations support this proposal, and Chapter 2 of this thesis provides evidence suggesting that in perspective taking people in fact attribute representations to others. Some supporting evidence comes from neuroimaging studies that found that brain regions that were found to respond selectively to belief-involving scenarios (Saxe & Kanwisher, 2003) does not differentiate between true and false belief scenarios (Young, Cushman, Hauser, & Saxe, 2007), which is consistent with the possibility that mental state calculation happens also in true belief attribution.

The relation between one's own representation and the others' attributed representation may also change within a situation. Consider the following scenario. Two people play (say, John and Tim) volleyball, and they both see the ball. Then the ball flies over a wall into another garden, and now neither of them see it; but if both have seen it fly behind the wall, both know it is there. They try to get the ball back by climbing over, but don't succeed; therefore John goes to get a ladder. However, while he is on his mission, someone from the other side of the wall sees the volleyball and throws it back. Tim sees this, and therefore now knows that the ball has been retrieved. When John returns with the ladder and approaches the wall, Tim will probably warn him that they don't need the ladder anymore. This will happen because Tim knows at this point that John has a false belief: he represents that the ball is behind the wall. But *when* did Tim attribute a representation of the ball to John? When John approached the wall? When he returned? When the ball was thrown

back? When the ball flew over the wall? Or even earlier, when they first saw the ball when they started playing?

Now consider another scenario. Anne and Mary studying together in the garden. Their mom brings out two cookies, and puts them on the table. Anne eats her cookie, but Mary wants to finish reading a chapter before she takes a break to eat hers. However, she is tempted by the sight of the cookie, so Anne wants to help her and puts the cookie behind a pile of books so Mary won't see it. Now Anne still sees the cookie from her side, but it is occluded from Mary, yet Mary knows that it is behind the pile of books. Mary continues to read, and at a certain point she puts her book down. Anne will now know, that Mary will reach behind the pile to get her cookie, based on her (true) belief that the cookie is there.

Many everyday scenarios are at least as complex as the ones described above, and in fact events often happen much faster, and may require fast reaction. It is therefore advantageous to track beliefs on-line, as the events are unfolding (Kovács, 2015). However, one probably does not track everyone's representation of *all* the things in the environment; but aspects of the environment need to become relevant or salient for them to be represented as the content of someone's belief (Carruthers, 2015a). Once for instance an object becomes relevant (such as the ball if we are playing volleyball), we start tracking others' epistemic access to it. When an event happens that changes the other's epistemic access to this object, we might represent this change, in order to then track her representation of it. Based on this reasoning, observing an object to become perceptually inaccessible from someone else's perspective, or observing someone seeing an object but not another one (a scenario

that is highly similar to e.g. Samson et al.'s 2010 task) might elicit the attribution of a representation to the other person.

It may turn out to be difficult to pinpoint the earliest stage when we attribute to someone a representation, or the smallest difference in epistemic access that elicits such computations. However, exploring these questions will allow us to gain a better understanding of the dynamics of the cognitive mechanisms that are involved in belief attribution.

Updating and discarding mental states

Automatic cognitive processes have been described with many different characteristics. One of the common attributes is that they draw only minimally on (or are independent of) attentional resources (Hasher & Zacks, 1979); and in some two-system theories of mindreading, related arguments are the basis of supposing a simple but fast system (Apperly & Butterfill, 2009), though this argument has been challenged by others (Schneider, Lam, et al., 2012). Automatic processes have also been claimed to take place under various circumstances in a constant manner (Hasher & Zacks, 1979). Relatedly, ToM processes were suggested to be always triggered by the presence of an agent (Leslie, German, & Polizzi, 2005), however, there is some evidence suggesting that it can depend on situational demands (Elekes et al., 2016). A yet unexplored aspect of automatic processes with regard to mentalizing is that such processes are often said to operate without a possibility to intervene (Shiffrin & Schneider, 1984).

In social interactions, there is often a need for updating or modifying one's own and others' representations. Therefore how much flexibility a mechanism responsible for mental state representations allows with regard to changes on the calculated representations, is of high relevance. However, most theories are rather vague with this respect. The minimal-ToM account suggests that the early operating system is fast and efficient, but inflexible (Apperly & Butterfill, 2009; Butterfill & Apperly, 2013). This claim mostly entails that it cannot compute certain mental state contents. But it is unclear how this system should react in case there is a need to modify the representation. Registrations are relations between an agent and an object. As such, they seem to be a relatively rigid structure. It is therefore likely that they are inflexible with regard to updates, and when there is a need for change, the entire representation needs to be calculated again. Leslie's view (Leslie et al., 2005) does not posit limits with regard to the scope of mental states the dedicated mechanisms can handle, nevertheless he assumes a modular functioning, which might be inconsistent with intervening in the process. On the other hand, the belief-file representations are suggested to allow for flexible changes of the elements, in rapid succession, without having to recomputed the entire representation (Kovács, 2015).

Consider an alternative ending to the Anne-Mary story above. After the cookie was put behind the books (which Mary saw, and might form the *true* belief that the cookie is behind the pile), while Mary is reading, their cat comes, walks around, and finds and eats the cookie. Anne sees the cat eating the cookie, but Mary doesn't notice a thing (and Anne doesn't tell her), therefore Mary now has a *false* belief that her cookie is behind the pile of books. When Mary puts down her book, Anne will know this means snack time; and will predict that Mary will reach behind the pile to get her cookie, based on her (false) belief; when in fact Anne knows the cookie is not there

anymore (in some sense it *doesn't exist*). During these events, when does Anne compute Mary's belief? When she reaches? When the cat eats the cookie (i.e. when Mary's belief turns false)? When Anne puts the cookie behind the pile? Or earlier?

Current theories lack a detailed account of the temporal dynamics and the flexibility of mindreading. There are many types of changes that may occur with regard to belief representations. One fundamental type of change that requires accessing the attributed representation may be when a person's representation becomes outdated, i.e. when a *true* belief becomes *false*. As we mentioned earlier, representing false beliefs does not have to entail representing the falsity itself. However, when a belief turns false, this might require differential processing, as from that on the other's representation has to be maintained as it is, while one's own might be updated. For example, in a change of location event, if the other person sees the initial hiding of an object, but not the relocation; her representation needs to be maintained, while one's own is updated with the new location. It is possible that in such situations one needs to access the attributed representation to strengthen it or to otherwise note that subsequent changes should not affect it. These predictions are currently untested.

On the other hand, there might be more radical changes that occur, and some changes may require revision rather than updating. One might be mistaken about some aspect of the situation; and occasionally we need to go back and overwrite our representations or inferences. For instance, we might misjudge whether someone really has seen something. On other occasions we might be mistaken on what they are focusing on. I may judge that my friend is checking her phone because she is nervous how a job interview went, but in fact she is waiting for her boyfriend to land and call

her; or she was looking at her phone thinking she should buy a new one. Even more radically, we may be mistaken about the agent whose intention we computed, such as when we misjudge a shadow for a person following us. Can such attributions be revised? Can mental state attribution be revoked, when the corresponding agent turns out *not* to be an agent in the first place? Leslie's theory predicts mental state computation to take place when there is an agent (Leslie et al., 2005); and other accounts also discuss automatic belief encoding triggered by the presence of an agent (Butterfill & Apperly, 2013; Kovács et al., 2010; Samson et al., 2010); however, they do not address whether once the process was set in motion, it could be stopped. If such computations happen in an automatic manner, they might be resistant to counterevidence once the representations have been formed. Chapter 5 in this thesis presents an empirical investigation of how infants can handle existing mental state attributions in light of novel information.

We set up this section to address some of the features of automatic processes with regard to Theory of Mind. However, we do acknowledge that some have argued that mindreading is not always automatic but sometimes spontaneous (i.e., independent from explicit but not from implicit goals; Carruthers, 2015a); or in some situations automatic and in others not (Butterfill & Apperly, 2013); or that mentalizing can be implicit but that not necessarily automatic by any strict criterion (Schneider, Slaughter, & Dux, 2015). While it is important to make clear the terminologies and the characteristics of the processes these entail; the issues mentioned here with regard to handling new information and incorporating it into existing belief representations, or revising belief attribution overall; are of importance regardless of the terminological distinctions.

Representing mental states are argued to be necessary for interpreting and predicting others' actions. Therefore it often needs to be fast and efficient. However, in social interactions rapid changes occur quite frequently, therefore it would be advantageous to track such changes on-line while the events are unfolding; and to react to them flexibly. In the present work (in Chapter 2 & 5) we aim to contribute to the characterization of how mindreading processes react to changes that necessitate the modification of existing mental state representations in light of new information.

1.3. Summary and outline of present work

Young infants' understanding of others' mental states has been documented in a wide array of tasks using spontaneous measures such as violation of expectation, predictive looking, pointing and helping behavior, and electrophysiological measures. However, the cognitive apparatus enabling Theory of Mind in infancy is a subject of debate. While as adults we can attribute any beliefs to others that we ourselves can entertain, in the case of infants it was recently suggested that they couldn't form certain kinds of representations as part of attributed beliefs. For instance, a suggested key limitation is that infants cannot encode that someone has a mistaken belief about the identity of an object (e.g. someone doesn't know that Clark Kent is Superman). Relatedly, it was proposed that even in adults such computations don't occur spontaneously but only in case of effortful deliberate mindreading.

Such considerations gave rise to accounts arguing that ToM undergoes a conceptual change between infancy and childhood, with some positing two separate systems responsible for mindreading: an early developing, efficient but inflexible; and a later developing, flexible but slow system. Others have criticized the necessity to

suppose distinct mechanisms for early and later mindreading; arguing that (i) development can be explained through a gradual development of both ToM-related and other cognitive abilities, and (ii) children's difficulties in certain situations, as well as adults' occasional lack of spontaneous mindreading can be explained by external factors such as the additional cognitive demands of the tasks. These claims are supported by findings suggesting a developmental continuity between early and later ToM abilities, and a common neural basis of spontaneous and elicited mindreading.

Based on these considerations, recent accounts reached back to some of the early work on mindreading that claim an innate basis of the capacity to represent others' mental states in the form of metarepresentations of propositional attitudes. Such proposals are modularist in nature to the extent that they suppose some early emerging specialized mechanisms dedicated to computing mental states. However, such a modular functioning is argued to be consistent with (i) development within the module, as well as (ii) a cooperative functioning of the modular part of ToM with other cognitive systems that themselves may undergo changes in development.

Early emergence accounts that posit a gradual development of ToM rather than conceptual change face two main challenges. First, they need to describe the cognitive processes that may underlie early ToM abilities. Second, predictions from alternative theories claiming conceptual change remain to be tested. The present work aims to contribute to these challenges, through addressing the cognitive processes that enable the developing mind to understand the social world around them. Describing what processes are present early on can lead to developmental models of Theory of

Mind that address the later additions and changes in these capacities; which in turn enables detecting if some capacities show an atypical development.

This thesis investigates the mental representations and cognitive processes underlying Theory of Mind in infants. It explores the possibility that ToM abilities in their core format are present from early on. On one hand, it aims to explore a possible mechanism through which belief contents are handled; where one's cognitive mechanisms that represent the environment are also recruited to compute representations attributed to others. This mechanism has three important implications. First, as it supposes a shared format of own and attributed representations, it explains well the modulation effect where ascribed representations influence one's own behavior (Chapters 3 & 4). Second, it predicts that signatures of representing primary representations should also accompany representing them as belief contents (Chapter 2). Third, it claims that infants should be able to handle belief representations with a variety of contents that they themselves can entertain (Chapter 3); and should not show arbitrary limitations with regard to the types of belief contents, but rather limitations that correspond to their own representational abilities (Chapter 4). In addition, there are many unexplored aspects of mindreading. Two of these were outlined in this thesis; the temporal dynamics of attributing belief representations (Chapter 2), and the flexibility of such processes when updating or revision is necessary (Chapter 5). The aim of the present thesis is to explore these questions to gain a better understanding of early ToM abilities.

Chapter 2 includes two experiments that address the neuro-cognitive bases of 8-month-olds' ability to encode the world from another person's perspective. We

measured gamma-band EEG activity over the temporal lobes, an established neural signature for sustained object representation after occlusion. We observed such gamma-band activity when an object was occluded from the infants' perspective, as well as when it was occluded only from the other person (Experiment 1), and also when subsequently the object disappeared but the person falsely believed the object to be present (Experiment 2). These findings suggest that the cognitive systems involved in representing the world from infants' own perspective are also recruited for encoding others' beliefs. In addition, these findings show that infants represent others' beliefs on-line, during the tracking of the unfolding of the events, and update these when a belief-relevant change happens.

Chapter 3 explores infants' abilities to attribute beliefs to others involving multiple objects. In a behavioral experiment with 14-month-old infants we assessed infants' manual search durations in various scenarios. We tested whether infants can ascribe to others beliefs regarding a potential hidden object, when this involved tracking the number of objects that were hidden and retrieved at a location (an opaque box) according to the other person's belief (Experiment 3). We showed scenarios where the other (falsely) believed an object to be present, or (falsely) believed the object to be absent. Results show that infants' search times were modulated by the others' belief. This suggests that infants were able to ascribe beliefs involving the tracking of multiple objects; and demonstrates the modulation effect in an active behavioral paradigm in infants.

Chapter 4 further explores infants' ability to ascribe beliefs involving objects. Using the same paradigm as in Chapter 3 we tested whether infants can attribute to others false beliefs involving the individuation and tracking of objects based on feature/kind

properties. Experiment 4 found the same modulation effect as Experiment 3 and 4, but here infants had to track the identity of objects to represent the other's belief about potential hidden objects in the box. Finally, Experiment 5 found that infants were able to attribute to another person a false belief based on *mistaken* individuation of two appearances of *one* single object as *two* separate objects. This latter finding demonstrated that infants can attribute to others beliefs about identity, in a spontaneous-measure paradigm.

Chapter 5 Explores the characteristics of infants' agency- and goal attributions. In a looking-time paradigm we tested whether (i) 9-month-old infants attribute goal-directedness to a self-propelled, animated box; and expect it to continue to act according to its goals (Experiment 6 & 7); and if they do, whether (ii) they can revise their expectations when they face evidence suggesting the box is in fact not self-propelled, therefore likely not an agent (Experiment 7). Infants indeed first seem to categorize the block as a goal-directed agent, and they potentially discard their expectation regarding the block's possible goal when given evidence on the lack of self-propelledness. These findings further characterize the flexibility of mental state attribution in young infants.

Chapter 6 summarizes the contribution of this work to the understanding of early socio-cognitive abilities. The findings are discussed in light of previous work, with recognizing possible limitations and outlining possible future directions.

Chapter 2

Neural correlates of representing others' beliefs in infants

"There is a gap between the mind and the world, and (as far as anybody knows) you need to posit internal representations if you are to have a hope of getting across it. Mind the gap. You'll regret it if you don't."

— Jerry A. Fodor

Representing attributed object representations

Human infants encode various aspects of the world, allowing them to successfully navigate their physical and social environment. In the introduction of this thesis we reviewed evidence suggesting that already in their first year of life infants can predict others' actions based on their mental states (Southgate & Vernetti, 2014), and by their second year this guides their active behavior (Buttelmann et al., 2009; Knudsen & Liszkowski, 2012a; Southgate, Chevallier, et al., 2010). A recurring debate is whether such findings can be taken as evidence that infants attribute beliefs to others and represent these belief contents in the form of metarepresentations (i.e., representations incorporating other representations) (Leslie, 1987). Some accounts question the validity of the interpretation of these studies in terms of mental state attributions, and suggest that instead of ascribing mental representations to others, infants simply store object-agent relations (Butterfill & Apperly, 2013), form associations, or apply behavioral rules (Perner & Ruffman, 2005). Similar alternatives were also raised with regard to nonhuman animals' Theory of Mind abilities (Povinelli & Vonk, 2003). Metarepresentations in general, and Theory of Mind or false belief understanding in particular, have been argued to be absent in other species than humans (Call, 2001; Suddendorf, 2012; Tomasello et al., 2003). Thus, to understand the nature and origins of such abilities it is a crucial question whether pre-linguistic creatures, specifically human infants attribute representations to other people.

Adult ToM postulates that an individual can attribute to others anything that the individual himself can represent. It is an open question whether this is the case in

infants (e.g. Apperly & Butterfill, 2009). In the introduction of the present thesis we suggested that one way to address this question is to look for similar cognitive processes at play in infants' reasoning about others' mental representations and in their own representational mechanisms. If infants store the representations they ascribe to others in a the same format as they store their own representations, these contents could be possibly as rich and complex as infants' own representations of the environment. In line with this, several theorists have proposed that the most efficient way to handle metarepresentations would be to re-use primary representations (Carruthers, 2013; Leslie, 1987; Sperber, 2000). We outlined two lines of research that provide evidence convergent with this proposal. On one hand, recent electrophysiological findings suggest that infants recruit their motor system not only when they perform an action but also during the prediction of others' actions (Southgate et al., 2009). Relatedly, research with adults showed that people use their cognitive mechanisms that are involved in detecting semantic mismatches from their own perspective, to detect such semantic mismatches attributed the confederate (S. Rueschemeyer, Gardner, & Stoner, 2014). On the other hand, behavioral evidence suggests that when infants or adults are exposed to situations where they can track others' perspective or beliefs, their own representations and the representations attributed to others seem to influence their reactions in analogous ways, which we referred to as the 'modulation effect' (Kovács et al., 2010; Samson et al., 2010).

In the experiments presented in this chapter we built on the proposal of re-using primary representations as the content of mental state representations, and hypothesized that if infants ascribe a representation to another person, say, about an object, they would rely on their original representation, which would then be used as the content of the mental state. This way infants' own representations of the

environment and the representations ascribed to others could be realized through one cognitive system subserving both processes. If so, this enables us to make predictions about the neural signatures of processing ascribed representations. For example, if maintaining a representation of an object as a primary representation has a specific neural correlate in infants, we should observe similar neural activation also if infants process an object representation they attribute to another person. To test these questions, we exploit earlier paradigms that found a specific brain signature accompanying object representations in infants.

Infants from a young age possess powerful representational abilities to sustain the representation of an object even if it is not visible to them anymore. Kaufman and colleagues (Kaufman et al., 2003) investigated the neural bases of sustained object representations, and found that gamma-band oscillatory activation in electroencephalographic (EEG) responses over the temporal regions increased when 6-month-old infants witnessed the occlusion of an object, compared to when the object disintegrated before occlusion. The authors argued that in the case of occlusion of the object infants have to maintain the object representation, because their visual access to it was compromised. Moreover, this activation did not simply reflect a memory trace of the object, since that should have also been present in the disintegration condition; rather it showed that infants actively maintained the representation of the object that they believed to be behind the occluder.

In another study Kaufman and colleagues (Kaufman et al., 2005) found similar activation when a toy train went behind an occluder, and then the occluder was lifted. Gamma activity was larger *before* the lifting of the occluder when according to infants' knowledge the train was still behind it, compared to a condition when they

saw it leave the scene before the lifting, hence expected the occluder to reveal an empty space with no train. This reflects that infants activated the object representation during the period before the lifting occurred. Following this, after lifting the occluder they varied whether infants saw an expected or an unexpected outcome. They found increased gamma-band activation when contrary to infants' expectations the train was absent when the occluder was lifted (unexpected disappearance event). The authors argued that in this case gamma activation reflected that infants had to deal with the mismatch between visual input and their object representation. Together, these results (Kaufman et al., 2003, 2005) suggest that in infants such gamma activation over temporal areas reflect accessing the representation of an object (such sustaining it or matching it to incongruent visual input). The experiments in this chapter put forward the hypothesis that such activation may not only reflect processes involved in how infants handle object representations for themselves, but also signal computations required for attributing a representation about an object to another person.

Another goal of the present study was to assess whether taking someone's perspective might involve ascribing a representation to them. Even though it might be impossible to decide whether we compute others' beliefs if their epistemic state is exactly the same as ours; it might bring us one step further to investigate representations of true beliefs if the perspective, and thus the underlying representation, is different from ours. Such a case of difference in perceptual access is when one agent sees an object while another does not. Would someone, who does not need to sustain the object representation (because the object is visible to him), represent another person's maintenance of that object (from whom it is occluded)? If

so, this could be evident in the above-mentioned indicators of sustained object representations.

In two experiments we presented 8-month-old infants with scenes involving an actor and an object, and recorded event-related EEG activity during events involving the occlusion of the object from the infants' or the actor's perspective. Experiment 1 tested whether in perspective-taking scenarios, when an object becomes occluded from another person's (but not the infant's) view, infants attribute to another person a sustained representation of this object. Experiment 2 addressed whether infants continue to represent this object when they see the object disintegrate, but the other person mistakenly believes it to still be present behind the occluder. An increase in gamma-band activation was predicted when either the infant, or the actor had to sustain the representation of the object.

2.1. Experiment 1

Experiment 1 explored 8-month-old infants' understanding of a scene where a person is attending to an object, which is then occluded from her. In order to test whether this event triggers an attribution process that involves sustained object representations, we developed scenarios involving occlusion events from multiple perspectives (see Figure 2.1). First, a target object and an actor were shown on the screen, with the object visible to both the infant and the actor. Then the object was occluded either from only the actor or also from the infant's view. In order to implement a dynamically changing visual access to the object from multiple viewpoints, we placed the object in a box that that had two sides removed. By rotating the box either the infant, the actor on the screen, neither, or both could see the object in question. We compared these events to scenarios where the box initially contained an object, but then the object disintegrated while both the actor and the infants could see this event. Therefore the motion of the box was identical in the two kinds of events, but in this latter case the box did not occlude an object from the actor's or infant's view, but rather just empty space.

We calculated the average EEG gamma-band activation (25-35 Hz) over the left and right posterior temporal regions specified by earlier studies targeting sustained object representations in infancy (Kaufman et al., 2003, 2005), during occlusion of the object from the actor's or the infant's view.

2.1.1 Materials and Methods

Participants

The final sample consisted of 15 full-term 8-month-old infants (mean age 8;6 [month; days]; age range 7;26-8;15). Twenty-four additional infants were tested but not included in the analysis due to fussiness (7), extensive body movements (4), insufficient number of trials (10), noise in the recording (2), or maternal interference (1).

Procedure

The same testing apparatus was used in both studies. Infants sat in a dimly lit soundproof room, 70 cm from a CRT monitor (100 Hz refresh rate), on their parent's lap. The parents were instructed to not to communicate with the infant. Infants watched a maximum total number of 60 trials. If infants were not attending to the screen and could not be reoriented, a short break was included. Videos of the infants were recorded during the presentation of the stimuli, in order to assess their looking behavior.

Stimuli

In each trial, a colorful fixation stimulus appeared on the screen (the fixation stimulus changed after every 10 trials), and a short sound was played before each trial to orient the infant's attention to the screen. The duration of the fixation stimulus varied randomly between 600-800 ms. If the infant did not orient to the screen, a

looming spiral appeared until they looked at the screen. During the trials, the videos were shown, each of them lasting 5300 to 5700 ms, depending on the length of the jittering periods.

The videos displayed a female actor and a rotating box that could contain one of 5 different objects varying in color and shape. Altogether 10 different videos were used, with 5 objects in each condition (Object Present - Occlusion vs. Object Absent - Occlusion), which were presented in a pseudo-random order with no more than 3 consecutive trials involving the same object or condition.

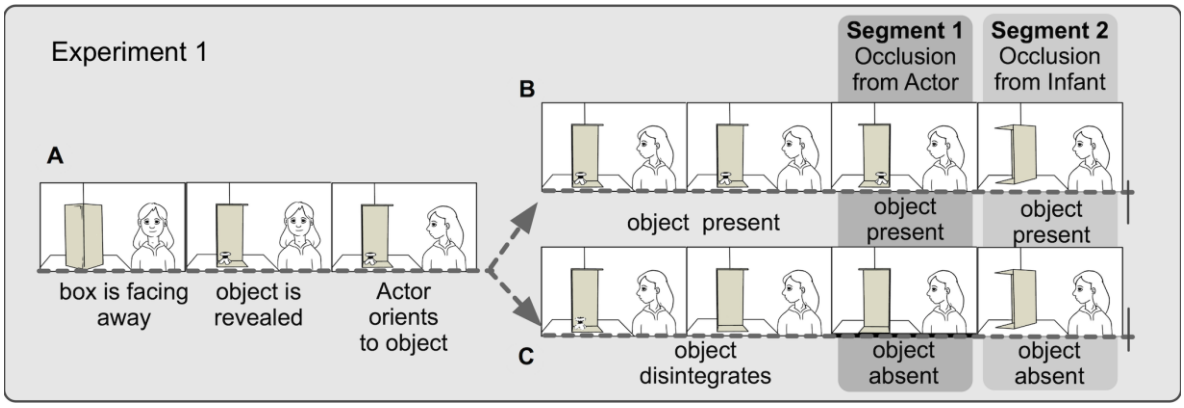

Figure 2.1. Schematic depiction of the events in Experiment 1. (a) The first 1.5 s of each video were identical in the two conditions. (b) object present—occlusion condition (c) object absent— occlusion condition.

Two types of videos were used, corresponding to two conditions. Both featured a female actor who looked at a rotating box open at two sides that contained an object. First, the opening of the box was facing away for 200 ms, then it rotated to reveal the object in 600 ms, and stood still for 200 ms. Then the Actor turned to the object in 600 ms. This was followed by the object remaining present (Object Present – Occlusion condition) or the object disintegrating in 600 ms (Object Absent – Occlusion condition). Following a 300-500 ms (randomized length) still period, the

box turned further, occluding the object (Object Present – Occlusion condition) or an empty area (Object Absent – Occlusion condition) from the Actor in 600 ms. After a 700-900 ms (randomized length) still period, the box rotated again further and occluded the object (Object Present – Occlusion condition) or an empty area (Object Absent – Occlusion condition) from the Infant as well. The trial ended with 800 ms still period with the box completely turned away (identical in the two conditions). The interval between trials lasted on average 1100 ms (randomized between 1000-1200 ms). In this interval first a blank screen, then the fixation stimulus was shown.

EEG recording and analysis

Continuous EEG was recorded using Hydrocel Geodesic Sensor Nets (Electrical Geodesics Inc., Eugene, OR, USA) from 124 channels equally distributed on the scalp, referenced to the vertex (Cz). The ground electrode was at the rear of the head (between Cz and Pz). The sampling rate was 500 Hz with a low-pass filter of 200 Hz. The EEG was segmented into two types of segments of interest.

Segmentation of continuous EEG data

The first segment (Occlusion from Actor) was the part of the video when, in the Object Present condition, the object was gradually hidden from the actor due to the rotation of the box, while the infants still saw it. In the Object Absent condition this segment included the identical movement of the empty box. This segment was time-locked to the start of the movement of the box, and lasted 1200 ms after rotation onset, of which the rotation took place in the first 600 ms. The second segment of interest (Occlusion from Infant) corresponded to the period when the object became

gradually hidden from the infants. This segment was time-locked to the start of the respective movement of the occluder and had a length of 1200 ms.

Transformed data were baseline-corrected to the average activity during the 200-ms-long baseline epoch. This epoch was the 200 ms recording preceding the rotation of the occluder in the Occlusion from Actor segment. In the Occlusion from Infant segment we used an epoch that roughly matched the baseline period in the first segment: a 200-ms-long interval ending 1500 ms before the onset of Occlusion from Infant (approximate comparison to baseline of Occlusion from Actor is due to a jitter period introduced between the two segments of a length varying between 100 and 300 ms). We selected this baseline based on the consideration that during this period the differences between conditions in terms of presence or absence of the object were already present, hence any possible difference in activation later cannot be simply due to the fact that in one condition infants saw an object during the crucial events whereas in the other they did not.

Artifact detection

Segments that were judged as not attended based on the video recording were excluded. Artifact detection and removal was performed using both automatic (by the Net Station tool, NetStation 4.4, Electrical Geodesics, Inc.), and manual methods. Segments with more than 20 channels (>~15% of the channels) containing artefacts (eye-movements, blinking, electrical noise) were excluded from the analysis. Bad channels were interpolated in the remaining segments. Infants contributed a mean number of 30 trials to the Occlusion from Actor segment (15/condition), and 24 to the Occlusion from Infant segment (12/condition). The lower number of average

trials in Occlusion from Infant was due to this segment being later in the trial, and infants became less attentive, resulting in a larger amount of artefacts. Infants who did not contribute a minimum of 10 trials per condition for the Occlusion from Actor segment (or 5 trials for the Occlusion from Infant segment) were excluded from the analysis.

The retained segments were imported into Matlab® using the toolbox EEGLAB (v9.0.5.6b) and re-referenced to the average reference. Time-frequency transformations were computed using EEGLAB and the custom-made script collection WTools (available at request) using continuous complex Morlet wavelets at each 1 Hz frequency bin between 5 and 60 Hz. An additional 400 ms of recording was left both at the beginning and at the end of the segments for the distortion caused by the wavelet transformation, which intervals were not included in the final analysis.

After the time-frequency transformation performed on the cleaned data, we compared oscillatory activity between the two conditions over 5-5 channels in right (channels 97, 98, 102, 103, 109, positioned above channel T3 in the 10-20 system) and left (channels 40, 41, 46, 47, 51, above channel T4 in the 10-20 system) temporal areas. Electrode sites were selected based on previous work by Kaufman and colleagues (16, 17). We analyzed the lower frequencies (25-35 Hz) of the gamma range, where activation was the most pronounced in earlier studies, (17) for our events of interest.

2.1.2 Results

First we analyzed gamma-band oscillatory activation in the two segments separately (see Figure 2.2), in two-way ANOVAs with Condition (Object Present - Occlusion vs. Object Absent - Occlusion) and Hemisphere (Left vs. Right) as within-subjects factors. To assess whether our results replicate earlier findings on neural signatures of sustained object representations, we analyzed activation during the Occlusion from Infant segment (Segment 2). Analyses revealed a significant main effect of Condition, $F(1,14) = 13.23$, $p = .003$, *partial* $\eta^2 = .49$, due to significantly higher activation in the Object Present - Occlusion ($M = 0.09$ μV, $SE = .03$), compared to Object Absent - Occlusion condition ($M = -0.07$ μV, $SE = .04$ μV, Figure 2.2B). There was no main effect of Hemisphere, and no interaction between Condition and Hemisphere ($F(1,14) = 0.04$, $p = .81$; and $F(1,14) = 0.06$, $p = .86$).

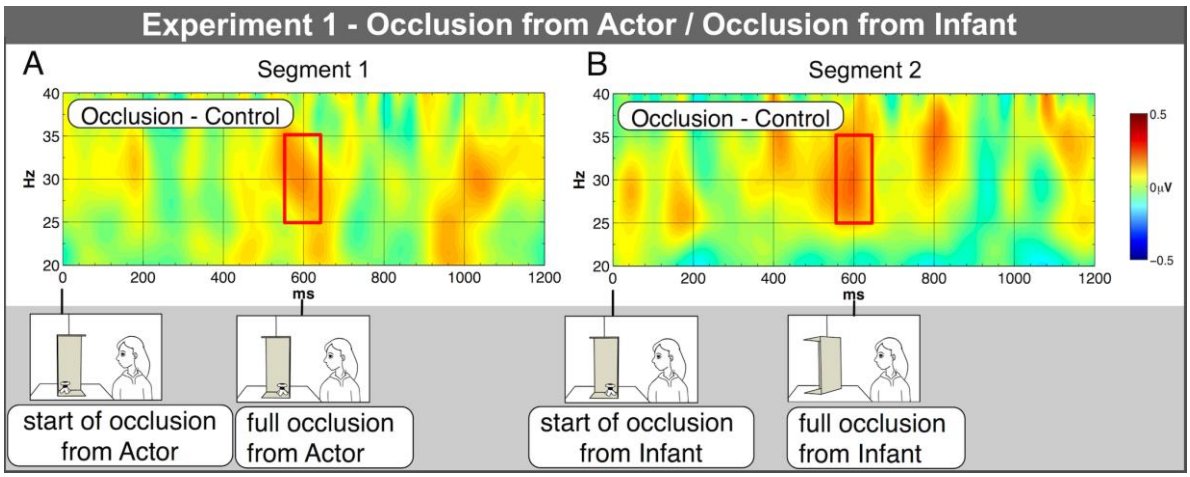

Figure 2.2. Time–frequency difference plots depicting average gamma-band oscillatory activation over the left and right posterior temporal cortex during the two segments in Experiment 1. Plots reflect mean activation difference between conditions; positive difference indicates higher activation in object present—occlusion condition than in object absent—occlusion condition. In both segments, 0 ms marks the onset of the occlusion event; in the first segment (a) from the actor, in the second segment (b) from the infant. Red rectangles indicate the time and frequency range over which statistics were computed.

We conducted a similar two-way ANOVA for the Occlusion from Actor segment (Segment 1), which revealed a significant interaction between Condition and Hemisphere ($F(1,14) = 4.99$, $p = .04$, *partial* $\eta^2 = .26$), and a marginally significant main effect of Condition ($F(1,14) = 4.53$, $p = .052$, *partial* $\eta^2 = .24$). There was no effect of Hemisphere ($F(1,14) = 0.06$, $p = .81$). To understand the interaction, we performed separate t-tests for the two hemispheres. There was no significant difference between Occlusion and Control in the right hemisphere, $t(14) = -1.03$, $p = .32$. Importantly, there was a significant difference in the left hemisphere, $t(14) = -2.56$, $p = .023$, $r^2 = .32$, due to higher gamma activation in the Object Present - Occlusion condition ($M = 0.08$ μV, $SE = 0.05$ μV) than in Object Absent - Occlusion condition ($M = -0.12$ μV, $SE = 0.04$ μV).

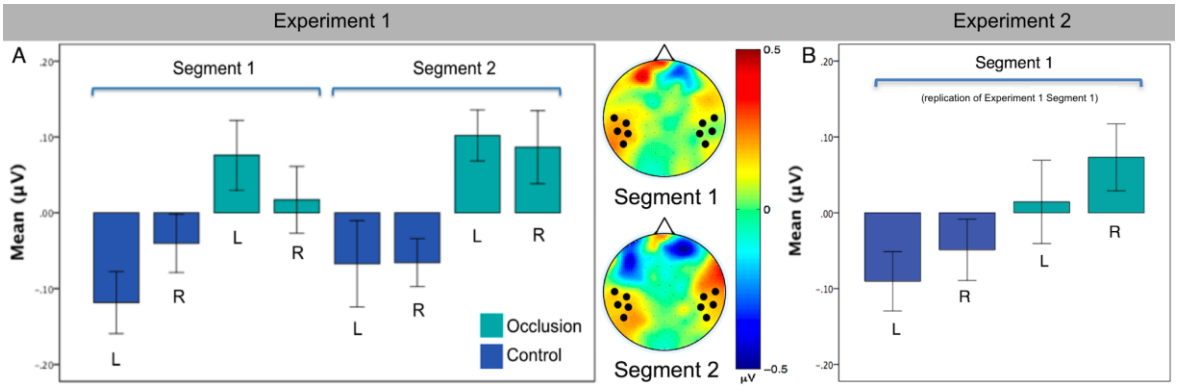

Figure 2.3. Mean activation in (A) Study 1 during Occlusion from Actor and Occlusion from Infant, and (B) Study 2 Occlusion from Actor at the target time windows (550-650 ms), at five left (L) and five right (R) temporal electrodes, over the 25-35 Hz frequency range. Error bars represent standard errors.

To assess whether the pattern of activation in the two segments was similar to each other, we analyzed them together in a repeated measure ANOVA with Segment (Occlusion from Actor vs. Occlusion from Infant), Condition (Object Present - Occlusion vs. Object Absent - Occlusion) and Hemisphere (Left vs. Right) as within-

subjects factors. We found a significant main effect of Condition, $F(1,14) = 13.24$, $p = .003$, *partial* $\eta^2 = .49$. No other main effect or interaction was significant (for mean values in Experiment 1, see Figure 2.3A). Thus, while in the Occlusion from Actor segment the effect was more pronounced on the left side, the direction of activation in this segment was similar in the two hemispheres and together they did not differ significantly from that in the Occlusion from Infant segment.

2.1.3 Discussion

Our results from the Occlusion from Infant segment are in line with earlier evidence pointing to a signature of infants' sustained object representation. Specifically, we observed higher gamma-band activation over posterior temporal areas when an object became occluded from the infants compared to when there was no object present (Occlusion from Infant). Crucially, we observed similar activation when the object became occluded from the actor only (Occlusion from Actor). Note that in the Occlusion from Actor segment the object was still visible to infants, therefore they did not have to sustain the object representation from their own perspective. This suggests that infants attributed a sustained representation of the object to the actor when she lost visual access to the object.

These results suggest that 8-month-old infants successfully computed the visual perspective of the actor regarding the object, an ability that is rarely observed at such a young age. Furthermore, while visual perspective taking (computing whether an agent can see an object) is necessary, it may not be sufficient to explain our findings. Taking the gamma-band oscillatory activity at the time of occlusion as an indicator of sustained object representation, infants in our study did not only infer that the

person no longer saw the object (as this would apply in the Object Absent - Occlusion condition as well); they also attributed to her the representation of the continued existence of the object behind the occluder.

Identifying the mechanisms at play when infants attribute a sustained object representation (a *true* belief) to another person allows further investigations of belief attribution processes. If the activation found in Experiment 1 accompanies events involving attributed object representations, then it should be present regardless of the veridicality of this representation, i.e., even when the other person holds a *false* belief regarding the object's existence behind the occluder.

2.2. Experiment 2

We developed a false belief scenario similar to the events in Experiment 1 (Figure 4). Eight-month-old infants were presented with the same initial event in which the actor attended to an object. Then in the critical condition the object became occluded from the person (Segment 1 – identical to Segment 1 Experiment 1), and afterwards the object disintegrated (Segment 2). This disintegration was therefore visible to the infants but not to the actor; hence this event must have resulted in the actor's false belief that the object was still behind the occluder. The critical question was whether infants would encode that the representation of the object cannot be discarded on behalf of the actor but it must be further sustained. Such an attribution process might be indicated by gamma-band activation during the disintegration event that is seen only by the infant but not the actor.

2.2.1 Materials and Methods

Participants

The final sample consisted of 15 full-term 8-month-old infants (mean age = 245 d; range = 229-261 d). Further 30 infants were tested but not included in the analysis due to fussiness (7), extensive body movements (10), insufficient number of trials (8), noise in the recording (3), experimental error (1), or maternal interference (1).

Stimuli

In Experiment 2, the setting of the scenes and the initial part of the videos (including the first segment of interest) was identical to Experiment 1. Then in the Object occluded – False Belief condition the object was occluded from the actor by the rotating box in 600 ms (Occlusion from Actor) and after a still period of 600 ms it disintegrated during 600 ms, while only the infants and not the actor could see this event (see Figure 2.4). We will refer to this disintegration period as False Belief event because in this case infants could note that the object ceased to exist and is not present anymore, and could infer that the actor should falsely believe it still to be present behind the occluding side of the box. In the Object Absent – True Belief condition the object disintegrated when the actor still saw the object, and subsequently the empty space was occluded in 600 ms. Following a 600 ms still period, (during the corresponding disintegration period of the Object Occluded – False Belief condition) in the Object Absent – True Belief condition the empty box remained turned away from the Actor for 600 ms. Thus, the two conditions differed only in the timing of the disintegration of the object: after (False Belief) or before (True Belief) it was occluded from the actor. Finally, in both conditions the empty box

rotated back towards the actor. Hence, infants in Experiment 2 never saw the object being occluded from them. The rotation of the box was identical in the two conditions.

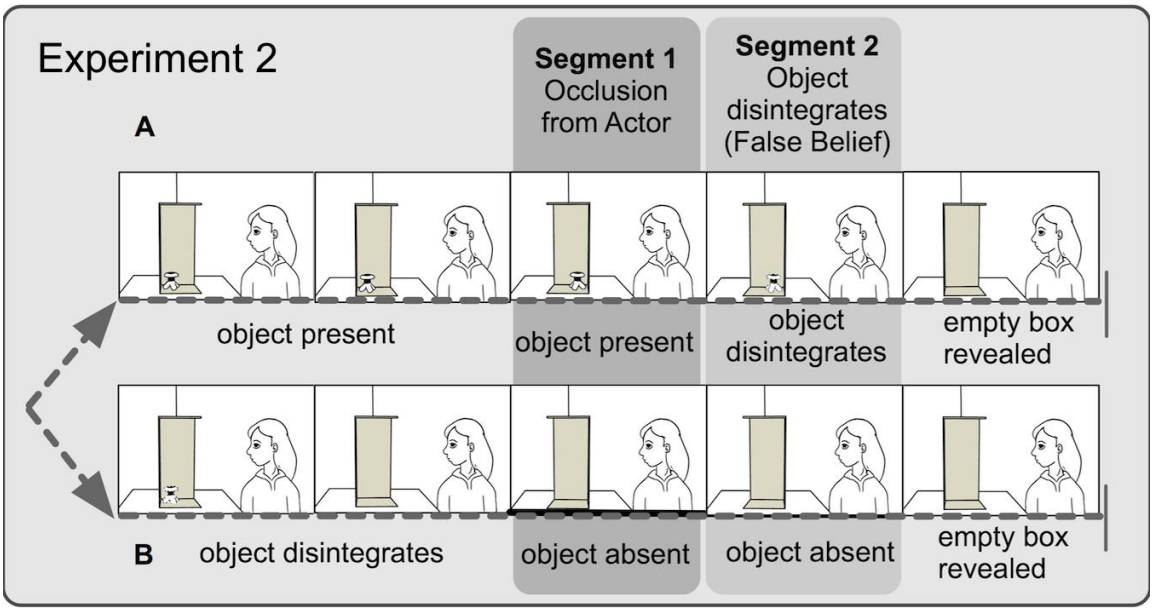

Figure 2.4. Schematic illustrations of the critical events in Experiment 2. The first 1.5 s of each video were identical in the two conditions (not depicted here). (A) In the object present—occlusion condition (B) Object Occluded – False Belief condition

EEG recording and analysis

Except for segmentation, EEG recording and analysis was identical to that of Experiment 1. Similarly to Experiment 1, the first segment (Occlusion from Actor) was the part of the video when the object was gradually hidden from the actor by the rotation of the box (in the Object Occluded – False Belief condition), while the infants still saw it; or the identical movement of the empty box (in the Object Absent – True Belief Condition). Hence, in the Occlusion from Actor segment, we specified the same time window of interest as in Experiment 1, and the baseline was again a 200-ms-long epoch finishing 1200 ms before the start of the segment.

The second segment of interest (False Belief Event) in Experiment 2 corresponded to the period when the object disintegrated after being occluded from the person (or the same time period during the Object Absent – True Belief Condition) and the subsequent still image. This period lasted 800 ms and its start was time locked to the start of disintegration event. Similarly to Experiment 1, the baseline was a 200-ms-long epoch finishing 1200 ms before the start of the segment (the same baseline as for Occlusion from Actor). In this False Belief segment, we analyzed activation throughout the disintegration event, from 1200 to 1800 ms, where 1200 ms corresponded to the onset of the disintegration and 1800 ms to the time point when the object had fully disappeared.

2.2.2. Results

We calculated the average gamma-band activation (25-35 Hz) the same way as in Experiment 1 during two Segments of interest: Occlusion from Actor and False Belief. As direct comparison between the two segments was not meaningful (one being an occlusion, which can be seen as a discrete event, while the other is a disintegration with a gradual temporal unfolding), activations in the two segments were analyzed separately. A two-way ANOVA on the Occlusion from Actor segment with Condition (Object Occluded – False Belief vs. Object Absent – True Belief) and Hemisphere (Left vs. Right) as within-subjects factors revealed a main effect of Condition ($F(1,14) = 5.98$, $p = .03$, *partial* $\eta^2 = .3$). This effect was due to higher activation in the Object Occluded – False Belief condition ($M = 0.044$ μV) than in Object Absent – True Belief ($M = -0.07$ μV, Figure 3B). No other main effect or interaction emerged.

We then compared activation during Occlusion from Actor in Experiment 2 to that of Experiment 1 (see Figure 2.3). These segments were identical in the two studies and both depicted an Occlusion from Actor event. A three-way mixed ANOVA was conducted with Condition (Object Present vs. Object Absent) and Hemisphere (Left vs. Right) as within-subjects factors and Experiment (1 vs. 2) as a between-subjects factor. This analysis revealed a main effect of Condition ($F(1,28) = 10.13$, $p = .004$, *partial* $\eta^2 = .27$), which was due to higher activation in the Object Present condition ($M = 0.05$ μV, $SE = 0.03$ μV) than in Object Absent condition ($M = -0.07$ μV, $SE = 0.02$ μV). There was no effect of Experiment ($F(1,14) = 0.01$, $p = .92$), and no interaction.

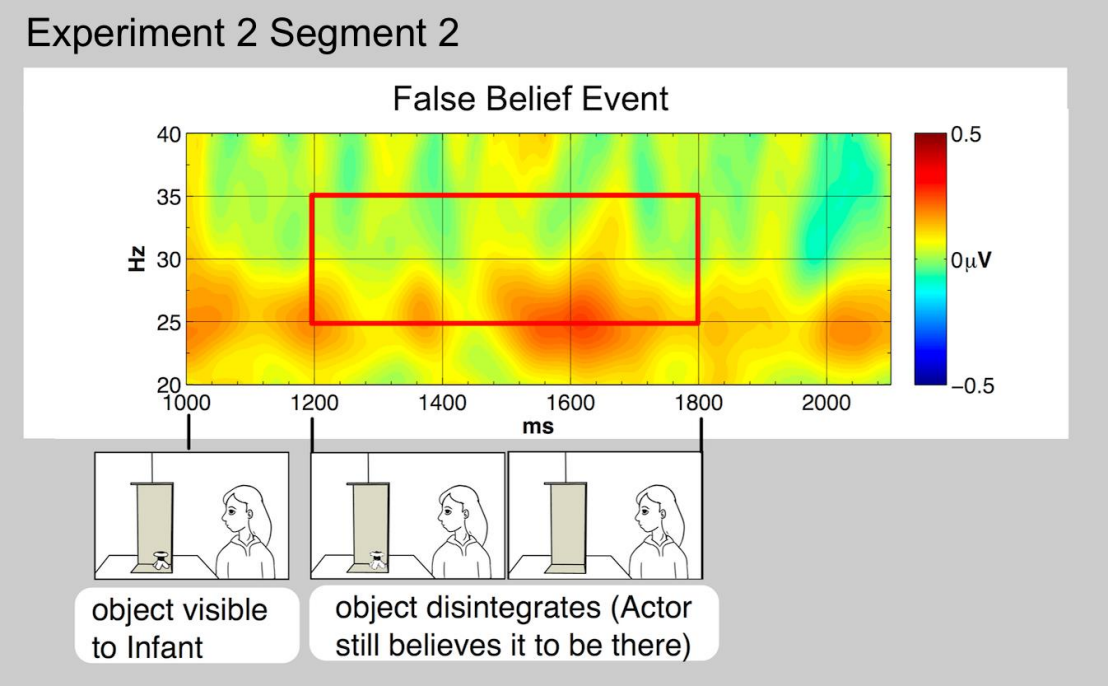

Figure 2.5. Time–frequency analysis of the average EEG during the false belief event at 10 electrodes over the left and right temporal cortex in study 2. The plot reflects mean activation difference between conditions; positive difference indicates higher activation in object occluded—false belief condition than in object absent—true belief. 1200 ms is the onset of the disintegration event and 1800 ms is the offset. The red rectangle indicates the time and frequency range over which statistics were computed.

Next we entered the activation during the False Belief segment of Experiment 2 in a two-way ANOVA with Condition (Object Occluded – False Belief vs. Object Absent – True Belief) and Hemisphere (Left vs. Right) as within-subjects factors. There was a significant main effect of Condition, $F(1,14) = 8.47$, $p = .01$, *partial* $\eta^2 = .38$, due to significantly higher activation in the Object Occluded – False Belief ($M = 0.07$ μV, $SE = 0.04$ μV), compared to Object Absent – True Belief condition ($M = -0.01$ μV, $SE = 0.05$ μV, Figure 2.5). There was no main effect of Hemisphere, and no interaction.

2.2.3 Additional Analyses (Experiment 1 & 2)

Late burst activation

In addition to analyzing activation in our predicted time windows, after visual inspection we analyzed activation in the Occlusion from Actor segment at a later time point, starting at 1000 ms after the onset of Segment 1. This interval was not part of our original time window of interest (time of full occlusion from the actor) as it was 400 ms after the object was completely occluded from the actor and there were no events happening at this time. Results are depicted in Figure 2.6. We analyzed this activation in Experiment 1 for the 1000-1100 ms period in a repeated measure ANOVA with Condition (Object Present - Occlusion vs. Object Absent - Occlusion) and Hemisphere (left vs. right) as within-subjects factors. We found a significant interaction between Condition and Hemisphere ($F(1,14) = 5.37$, $p = .036$, $\eta^2 = .277$). We found a marginally significant effect of Condition in the left hemisphere ($t(14) = -2.07$, $p = .057$, $r^2 = .23$) with higher activation in the Object Present - Occlusion condition than in the Object Absent - Occlusion condition ($M = 0.09$ μV, $SE = .05$ μV, and $M = -0.11$ μV, $SE = .07$ μV, respectively). There was no effect of Condition in the

right hemisphere ($t(14)$ = -0.42, p = .68). A similar analysis did not yield any significant main effects or interactions in Segment 2. Since this late activation burst in Occlusion from Actor segment was not expected, we intended to replicate it to assess the robustness of the finding in Experiment 2.

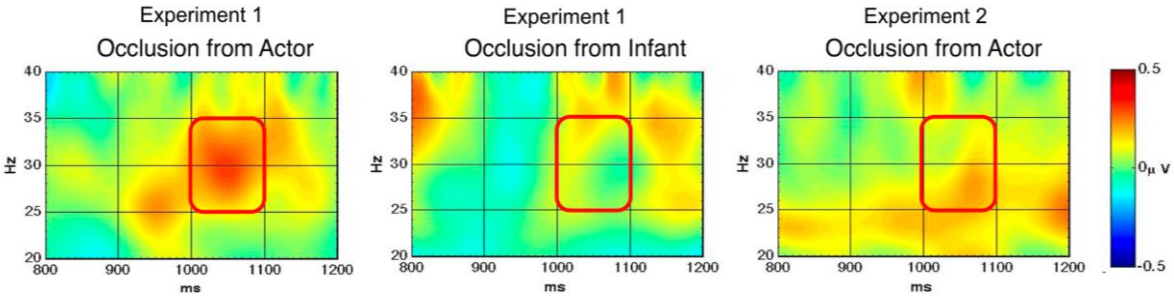

Figure 2.6. Time-frequency difference plots depicting average gamma-band oscillatory activation over the left and right posterior temporal cortex showing the late burst activation. Plots show 800-1200 ms after segment onset and reflect mean activation difference between conditions; positive difference indicates higher activation in Object Present condition than in Object Absent condition. Red rectangles indicate the time and frequency range over which statistics were computed.

After performing Experiment 2 we analyzed the late burst activation from the two studies in the Occlusion from Actor segment in a 3-way mixed ANOVA with Condition (Object Present vs. Object Absent) and Hemisphere (Left vs. Right) as within-subjects, and Experiment as between-subjects factor. There was a significant main effect of Condition ($F(1,28)$ = 5.28, p = .029, *partial* η^2 = .16), due to significantly higher activation in the Object Present condition than in Object Absent condition (M = 0.05 μV, SE= 0.02 μV, and M = -0.05 μV, SE= 0.03 μV, respectively). There was also a significant Condition*Hemisphere interaction ($F(1,28)$ = 4.82, p = .037, *partial* η^2 = .15). We analyzed activation separately in the two hemispheres in a two-way mixed ANOVA with Condition (Object Present vs. Object Absent) as within-subjects factor and Experiment (1 vs. 2) as a between-subjects factor. On the left hemisphere there was a significant main effect of Condition ($F(1,28)$ = 7.97, p = .01, *partial* η^2 = .22), due

to higher activation in the Object Present (*M* = 0.07 μV, *SE* = .03 μV), than in Object Absent condition (*M* = -0.09 μV, *SE* = .04 μV). Similar analysis on the right hemisphere did not yield any significant main effects or interactions.

This additional burst of activation therefore was present in both studies towards the end of the Occlusion from Actor segment. At this time nothing was happening in the video, therefore this activation likely reflects computational processes that involve further processing of the earlier observed events. Since this late activation accompanied only processing the object occlusion from the other person, it might reflect some further processing or tagging the sustained representation as ascribed to the other person, hence possibly playing a role in distinguishing an ascribed representation from the infants' own reality representation.

Ruling out potential ocular artifacts

Recently an objection was put forward regarding the interpretation of scalp-recorded gamma-band EEG activity in adults as a correlate of object processing. Yuval-Greenberg and colleagues (Yuval-Greenberg, Tomer, Keren, Nelken, & Deouell, 2008) reported that in human adults saccadic spike potentials (SP), co-occurring with micro-saccades (MS), contribute to this signal, and questioned the neural origins of the oscillatory activation found in earlier studies. In response to this, specific tools have been developed to remove possible MS-related artifacts from adult EEG data (e.g., (Hassler, Barreto, & Gruber, 2011). Köster (2016) pointed out that analogous attempts have not been implemented in infancy research. While this is indeed the case, there are several theoretical and methodological considerations that cast doubt on whether it is necessary or possible to apply these tools to infant EEG recordings.

First, the algorithm applied on adult EEG to remove MS-related artifacts would not be applicable to infant recordings as it is. Hassler *et al.* (Hassler et al., 2011) propose a two-step method, which consists of detecting and then removing SPs that accompany MSs. The first step of this method detects SPs based on their characteristics in adult EEG. However, Csibra and colleagues (Csibra, Tucker, Volein, & Johnson, 2000) found no saccade-related SPs in infants younger than 12 months, and even at this age SPs differed greatly in amplitude and in morphology from those reported in adults. Because of this, the algorithms used with adults to detect SPs would simply not be applicable to infant EEG. The second step of Hassler *et al.* (2011), using independent component analysis (ICA) to remove MS-related SPs from the signal, also seems unfeasible to apply directly on infant data given the nature of infant EEG recordings. As Köster (2016) rightly points out, performing ICA requires a vast amount of data to produce valid results. As an estimate, finding N stable components in N−channel data requires more than 3*N^2 sample points at each channel ("EEGLAB Tutorial", 2001). In EEG recordings at 128 channels and 500 Hz sampling rate (like in our study) this requirement demands more than 90 seconds of perfectly clean EEG on *all* channels. In most infant EEG studies (especially ones with relatively longer trials and dynamic stimuli), movement artifacts regularly contaminate recordings, and the cleaned data are much sparser than what might be required by ICA.

Furthermore, to our knowledge no one has managed to identify and measure MSs in infants so far, and therefore it is not known in what form they occur at this early age. While the appropriate tools are available (eye-trackers with a high enough sampling rate), it would be a separate methodological challenge to keep young

infants' head sufficiently stable for accurately measuring MSs. Therefore even in case of successful co-recording of EEG and eye-movements, it is unclear how MSs (and/or SPs) should be detected. Because of this, at the moment it is not possible to remove any potential MS-related artifacts from infant EEG, and we agree with Köster (2016) that we cannot decisively exclude the possibility that microsaccades contaminate gamma-band responses in infants.

To estimate the likelihood of eye-movement contamination of our measures in our data, we performed an additional analysis on our time-frequency data from Study 1. To approximate a measure of eye-movement-related activity, we estimated the bipolar horizontal EOG signal in our recordings by subtracting the activation at the two electrodes closest to the outer canthi of the eyes (channels 1 and 32) from each other. We then subjected this signal to the same time-frequency analysis as our original data (for a visualization of results see Figure 2.7.), and correlated the resulted gamma activation in this EOG signal with the activation we obtained in our original analyses. If eye-movements induced the gamma-band activation found in our study, then activations at the temporal channels would likely be correlated with the EOG signal. However, this correlation was not significant either in Segment 1 (r= .347, p=.205 in Occlusion condition - for activations in the Occlusion condition during Segment 1, see Figure 7; and r= .239, p=.390 in Control condition), or in Segment 2 (r= -.059, p=.835 in Occlusion condition, and r= .-099, p=.725 in Control condition). Based on this analysis it seems unlikely that our findings originate from eye-movements.

Additionally, beyond the methodological challenge to detect MS-related artifacts in infant EEG, several findings (including some mentioned by Köster, 2016) of scalp

recorded gamma-band activity during object processing in infants would not be easily explained by MS patterns. First, in many cases, there were no visual differences during the measurement periods between the experimental and control conditions, and therefore it is not clear why MSs would show a different pattern (e.g. Kaufman et al., 2003; Leung et al., 2016). Second, many of the studies reported gamma-band activity over temporal areas (e.g. Kampis, Parise, Csibra, & Kovács, 2015; Kaufman et al., 2003), whereas MS-related SPs were found mostly around the midline in adults (Yuval-Greenberg et al., 2008). Third, while MS-related SPs were shown to manifest themselves in a time window of approximately 200-350 ms after stimulus onset, many studies have used different time windows for analyses (Kampis et al., 2015; Kaufman et al., 2005), and in some cases it is not obvious what should count as stimulus onset, as activation was measured after a longer sequence of events (Kampis et al., 2015; Kaufman et al., 2003). Finally, as Melloni and colleagues (Melloni, Schwiedrzik, Wibral, Rodriguez, & Singer, 2009) pointed out in their response to the paper demonstrating MS-related gamma activity, MS-related EEG effects should show a broadband response, whereas many studies report effects in narrower gamma ranges, and this observation applies to infant recordings as well.

In sum, on one hand the tools developed for MS-related artifact removal from adult EEG are not used currently in infant EEG because they are not straightforwardly applicable to infant data. Once our understanding of the characteristics of infant EEG and (oculo-)motor development reaches the necessary level, it will be possible to return to these concerns and address them.

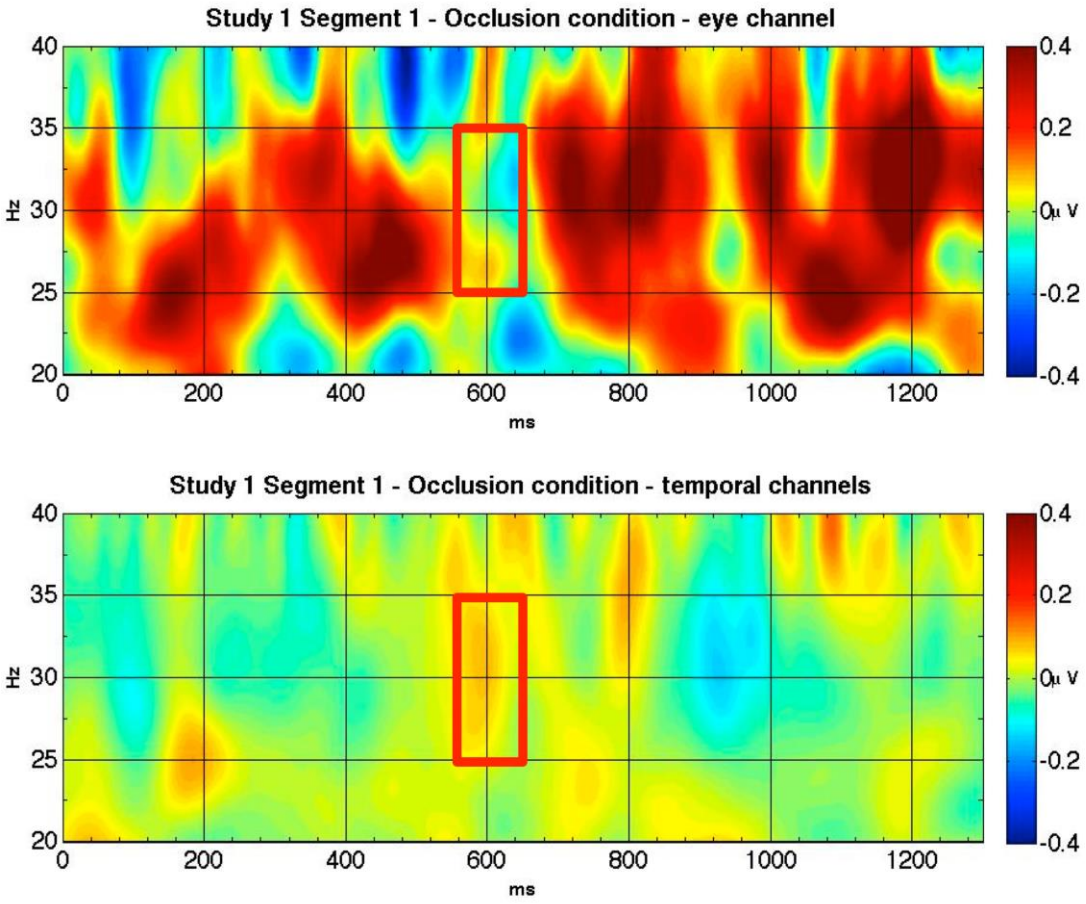

Figure 2.7. Gamma-band activation in the eye channel (channel 32 subtracted from channel 1), and temporal channels (41, 41, 46, 47, 51, 97, 98, 102, 103, 109). The red rectangle marks the frequency and time window used in the main analyses.

On the other hand it is not clear whether this issue has to be addressed in infants, as the factors that were found to induce possible artifacts in adult studies are not simply hard to measure but might not be present (or might have radically different characteristics) in young infants. With regard to our own data, it seems unlikely that the gamma-band activation in temporal areas was due to infants' eye-movements during the observation of the events (Figure 2.7). Finally, some recent results, discussed in Köster (Köster, 2016) as well, suggest that gamma-band oscillations, even in the adult literature, provide us with a valid tool to investigate object representations (Köster, Friese, Schöne, Trujillo-Barreto, & Gruber, 2014).

2.2.4. Discussion

The results of Experiment 2 are consistent with the proposal that infants ascribe object representations to others not only when they attribute true beliefs, but also when they can attribute false beliefs to them. Similarly to Experiment 1, belief attribution here was based on visual perspective taking (infants had to encode that the object was not visible to the person). Crucially, in the False Belief segment, when the object disintegrated and this was visible to the infant but not to the person, there was increased gamma-band activation, similarly to the occlusion events (occlusion from the infant or from the person).

These results suggest that infants encode that the other person continues to represent the object, despite evidence that prompts them to discard their own representation of the very same object. Since disintegration has been previously shown not to trigger sustained object representation (Kaufman et al., 2005), higher gamma activation during this event reflects that the infants sustained the object representation they had attributed to the actor (who falsely believed the object to be behind the occluder), even though this representation was in sharp conflict with the infants' own perception (as the object disintegrated). Thus, the infants must have encoded that the other person had seen the object being occluded, but did not see the disintegration, and hence the attributed object representation could not be discarded on her behalf, but had to be possibly refreshed and sustained further.

We see no obvious ways to explain the activation patterns we observed in Experiment 1 and 2 in terms of simpler cognitive mechanisms that do not involve belief attributions. First, activation during occlusion from the actor only (Occlusion

from Actor segments in both studies) could not be due to infants' own sustained representation, since they continued to see the object during this event. Second, our results cannot be attributed to perceptual differences between the conditions (e.g., that the object was present in one condition but not in the other), since we subtracted the corresponding baseline activation from our data where this difference already existed, hence any activation difference due to this factor would have been thus subtracted from the time window of interest. Furthermore, results from the Occlusion from Actor segment in Experiment 2 excluded the possibility that the gamma-band activation in the Occlusion from Actor segment was due to infants' expectation of occlusion from their own perspective, as no such occlusion followed.

Additionally, results from Experiment 2 confirm the late burst activation we found in Experiment 1. This additional burst of activation therefore was present in both studies towards the end of the Occlusion from Actor segment. During this period that followed after the occlusion of the object from the actor, nothing was happening in the video. Therefore this activation likely reflects computational processes that involve further processing of the earlier observed events, possibly related to keeping in mind the object representation attributed to the actor.

2.3. General Discussion

The goal of the present chapter was to investigate whether young infants ascribe representations to others during tracking of what another person sees, knows or believes, through utilizing their own representational system that is otherwise used for encoding objects and events in the world. In Experiment 1 we presented infants with scenes depicting a simple situation involving an object and actor, and events

where the infant's or the other person's perceptual access to the object changed dynamically. In Experiment 2 we constructed a case where this event could lead to a false belief about the presence of the object in the other person. We recorded event-related oscillatory activity during the observation of these events.

Earlier studies (Kaufman et al., 2003, 2005) found gamma-band oscillatory activity in infants for sustained object representation. We found similar gamma-band activations when an object became occluded from the infants' own (Experiment 1, Occlusion from Infant) or someone else's perspective (Experiment 1 & 2, Occlusion from Actor), consistent with the possibility that there are shared underlying mechanisms for sustained object representations for the self and for the ones attributed to another person. Crucially, the activation found in response to object occlusion from the other person's perspective could only be explained by the enrolment of an object representation ascribed to her. This is supported by the fact that during this interval infants continued to perceive the object and therefore did not need to sustain the representation for them. Importantly, the same activation was observed in a false belief situation where, after being occluded from the actor, the infant saw the object disintegrating (Experiment 2, False Belief segment). Due to disintegration the object ceased to exist from the infant's point of view, therefore EEG activation during this event is likely due to a sustained object representation on behalf of the actor. Together, the activations we found are indicative of the on-line processing of a representation that infants attribute to another person – a metarepresentation - based on her earlier perceptual access.

The finding that the cognitive systems that are otherwise dedicated for representing objects are also involved in mentalizing processes points to the

possibility that infants recruit cognitive systems from outside of a hypothesized ToM-network (Saxe & Kanwisher, 2003) or ToM module (Leslie, 1994) when representing others' beliefs. Yet, we do not take such data to speak to the question that has repeatedly emerged with regard to ToM capacities, namely, whether such reasoning is predominantly subserved by domain-general or domain-specific processes (Leslie & Thaiss, 1992). The gamma activations found in the 'Occlusion from Actor' events most likely signal sustaining an attributed representation of an object. This process relates to the encoding of the *content* of the actor's belief, in other terms to the formation of a metarepresentation of this belief content. However, as this is likely one of the first steps in the process of belief ascription (Kovács, 2015), our findings leave open the possibility that in the further steps of belief processing such representations would serve as input to more specialized mindreading processes.

Together, these experiments provide electrophysiological evidence suggesting that preverbal infants engage in encoding the visual perspective and the false belief of others. By possessing such powerful representational capacities, infants may be endowed with the ability to ascribe to others any representations they themselves can form, including representations that are in conflict with their own representation of reality. In Chapters 3 & 4 we present a series of experiments that explored infants' abilities to represent others' mental states with a variety of contents.

Chapter 3

Attribution of beliefs involving tracking more than one object

3.1 Introduction

In Chapter 2 we discussed the very foundations of what mechanisms enable young infants to represent other people's mental representations. We examined the highly prevalent case of tracking objects in the environment, and representing from another person's point of view how this object is encoded. We presented evidence that is consistent with the possibility that even infants as young as 8 months of age can use their cognitive apparatus underlying their primary object representations, to meta-represent an object as the content of someone else's (true or false) belief. As discussed in the introduction, in representational terms, attributed beliefs involving tracking objects could be encoded with representations that build on the resources of the object-tracking system. As from first-person perspective the object-tracking system is capable of tracking multiple objects, it seems reasonable to assume that the belief reasoning processes could build on this and would be capable of representing beliefs that require tracking more than one object. We first characterize some mechanisms of object representations from infants' first-person perspective, and then discuss how these could be involved in belief representations.

Tracking multiple objects from first-person perspective

In order to successfully navigate in the world, infants need to master extensive knowledge about physical objects and events. In the introduction we discussed the involvement of object-files and indexes in encoding and tracking objects in the environment (Pylyshyn, 2001). Object-files are mid-level object representations that are linked to objects through the indexes, and thus enable tracking objects continuously through space and time, such as during brief occlusion of the object

(Scholl & Pylyshyn, 1999). Through assignment of distinct symbols (files) to each object, information stored about the objects in this system also implicitly represents information about the *number* of objects that are tracked at any given time, and therefore numerical equivalence between two arrays can be established through one-to-one correspondence of the files. This system was therefore suggested to also be involved in quantifications involving small object arrays, up to a set-size signature of around 3 items in infants (Feigenson & Carey, 2005; but see Cordes & Brannon, 2008).

Quantification in some cases requires representing items of an array after the items become invisible, and tracking the items of such arrays constitutes the basis of simple mathematical operations such as addition or subtraction. Infants were found to be able to perform simple arithmetic operations on small number of items already around 5 months of age (Wynn, 1992). Infants can individuate and track multiple objects at a time, and through their first year of life they can store more and more information about these objects (Leslie et al., 1998; Wilcox, 1999). Around their first birthday, when their motor abilities enable them, they can use these abilities to actively make choices and explore the environment (Feigenson & Carey, 2003; Feigenson et al., 2002).

Feigenson and Carey (2003) used a manual search task (by Van de Walle, Carey, & Prevor, 2000) with 14-month-olds, where they presented infants an opaque box, and then placed a certain amount (2, 3 or 4) of objects inside. Next, infants were allowed to retrieve these objects. Measuring the duration of infants' manual search in the box showed that infants keep searching for an object if they have not yet retrieved all the objects that they previously saw being placed inside. Specifically, infants

searched longer in the box if one of the objects was still inside, compared to when all objects that were placed in the box were already retrieved. These findings suggest that infants are able to represent and track multiple objects, and these representations guide their active behavior. Such abilities are essential already at a young age, as in everyday life there are plenty of objects in the infants' environment. Storing information about them, and tracking them through space and time enables manipulating them, and learning about them. This, in turn, enables infants to acquire knowledge, among others, on human material culture.

Tracking multiple entities in belief representations

The majority of social interactions in humans', and in particular in infants' lives, involve objects around them. Therefore it is not just essential that infants themselves can track objects in the environment, but it is also necessary for them to know what others around them know or believe about these objects. Such representations are crucial for learning about a variety of domains like tool use or language.

Infants' abilities to track beliefs involving multiple entities around them are currently unexplored. For instance, little is known what happens once we increase the number of factors to keep in mind from the one person-one object scenario; such as the number of agents, or the number of objects involved. Some studies have used more than one object, but one might argue that these still only required focusing on one object, and *ignoring* another one. Southgate, Chevallier, & Csibra (2010) showed 17-month-old infants that an agent witnessed two objects being put respectively into different boxes. Then in the False Belief condition the objects were swapped between

the boxes, while the agent was turned away. When she turned back, she named the object in one of the boxes (e.g. sefo), and then asked the infant to give her the sefo. In this False Belief condition, infants tended to give her the object that she believed to be in the box, rather than the object that was really in the box. Infants in this case therefore demonstrated an understanding of her belief regarding which object is in that box (or in other words, to track where her referred object is). However, to succeed on this task, infants did not have to track two objects from the other person's perspective, but only the other person's belief about the location of one object (i.e. the one she requested).

To our knowledge there are only two studies that go beyond testing infants' and children's capacity to represent *one* agent's belief about *one* object's location, and looked at how children's working memory supports tracking others' beliefs. In one (Wang & Leslie, 2013) they tested whether 2-year old children can keep track of two agents' beliefs in an anticipatory looking paradigm. Children successfully anticipated the agents' actions based on their respective beliefs, suggesting that at the age of 2 years children can keep track of at least two separate agents and correctly bind two distinct beliefs to these agents.

In another study Cheng and colleagues (Cheng, Wang, & Leslie, 2016) tested whether 3 and 4-year-olds can track False Beliefs of multiple agents in a modified change-of-location task. The task started as the often-used Sally-Anne task (S Baron-Cohen et al., 1985): a protagonist 1 hides her toy in location A, then in her absence another actor (protagonist 2) moves the toy from location A to B. In the traditional task the event stops here and children are asked where the protagonist 1 thinks the toy is. However, in this study they continued the event along a similar logic, and

brought in protagonist 3, who in the absence of protagonist 1 and 2 moved the toy to location C. This sequence then continued up to four different agents with corresponding false beliefs (protagonist 1 – box A, protagonist 2 – box B, and so on). Finally, children were tested in two kinds of trials: a low-demand and a high-demand. In low-demand the toy was removed from the scene in the end, and therefore when children were tested on the beliefs of protagonists 1-4, the toy was absent from any of the boxes. In the high demand the toy remained in the last box. They found that in the low-demand condition even 3-year-olds successfully tracked the False Belief of up to three agents (but performance broke down at 4 agents); and 4-year-olds could keep track of the False Beliefs up to at least 4 agents. In the high-demand condition 3-year-olds showed a True Belief bias (as in traditional tasks), and 4-year-olds showed correct performance up to 3 agents but not with 4 (where their performance resembled that of 3-year-olds').

Together, these two studies provide evidence that children at least from the age of 2 years can sustain more than one belief representation in parallel. However, both of these cases involved *multiple agents*, and separate beliefs of these agents about *the same object*. It is therefore an open question whether infants can keep track of *more than one object* as the content of *one person's* belief.

We aimed to test whether infants can track more than one object as someone else's belief content and perform simple calculations with these contents. We intended to use a continuous measure as this proved to be an effective assessment of such cognitive computations from infants' own perspective (Feigenson & Carey, 2003; Zosh & Feigenson, 2012). However, it is unclear whether we should expect infants to predict others to search longer in cases where they themselves would

search longer, as infants are unlikely to reflect on these subtle differences in their duration of search, and therefore it is not clear what predictions they would make in such cases. We reasoned that finding an indirect measure might provide better insight into infants' representations. Infants' (and adults') behavior seems to be modulated in some cases by another agent's belief contents (Kovács et al., 2010; Samson et al., 2010). Similarly, infants' manual search after observing objects being placed into a box might also be influenced by another protagonist's beliefs about the number of objects in that box. To support this rationale, we first refer to the phenomenon we introduced in the introduction, the modulation of one's behavior by others' beliefs; which provides the basis of our measure.

Continuous measures grasp modulation of one's own behavior by others' beliefs

Tracking others' beliefs enables predicting others' actions, interpreting them, and reacting accordingly[14] (Liszkowski, 2013). For this, one needs to accurately separate attributed beliefs (that can be used to predict others' actions) from one's own knowledge (used to plan one's own actions). However, as supported by findings reported in Chapter 2 of the present thesis, one's own representations and attributed belief contents likely draw on common resources and share their representational formats. Therefore tracking others' beliefs results in holding multiple representations that are highly overlapping in their content and format (i.e. one's own representations, and attributed ones). Keeping these separate might be resource

[14] We mean 'action' in a rather broad sense, incorporating for instance, communicative acts as well.

demanding and might not always be executed properly. This is in line with the growing body of evidence (see Chapter 1.2.2. of the present thesis) showing that attributed beliefs occasionally intrude in one's own representations and influence one's behavior through modulating responses.

Continuous measures, such as reaction time, were found to be modulated by various factors, for example cognitive interference (Stroop, 1936), semantic priming (Becker, 1979), and as discussed before, other people's perspective or beliefs (Kovács et al., 2010; Samson et al., 2010). We reasoned therefore that manual search duration could also possibly be a good candidate to be modulated by such factors. Crucially, in the manual search task when infants' search duration is measured, they never successfully retrieve any objects, as even when there should be some in the box, they are in fact hidden in a secret compartment. This is done in order to avoid search duration to be determined by when the infant happens to stumble upon the object. Therefore this task provides a continuous measure of how inclined infants are to search further. While categorical answers might be too resilient to be altered, a small but consistent shift in duration on a continuous measure such as search duration could be a more sensitive measure of the influence of external factors (such as another person's representation about a task-relevant aspect of the environment) on behavior.

3.2. Experiment 3: representing others' beliefs about multiple objects

Earlier we suggested that infants exploit their resources dedicated to represent the environment in the service of mindreading to represent others' belief contents. If infants are indeed able to make use of their object-tracking system to represent others' beliefs, they should be capable of tracking multiple objects in false belief scenarios. Similarly to the simple arithmetic calculations they proved to be capable of from their own perspective in Feigenson & Carey's (2003) studies, infants should be able to perform such calculations on someone else's belief contents.

In order to investigate these questions, we developed a paradigm based on the manual search task of Feigenson and Carey (2003). We designed scenarios in pairs, which were matched with regard to what the infants know about the content of the box (i.e. always empty or always containing one object), and manipulated what the other person believes to be in the box. We assessed the duration of infants' manual search in these different scenarios, and compared search duration within the scenario pairs. Infants were previously found to search longer when they themselves thought there was still an object remaining in a box. If infants' reactions can be modulated by the other person's belief, in our case this could manifest in longer search times when not the infant, but only the *other person* believes that an object still remains in the box. Since the only difference between the scene pairs was the other person's belief regarding the content of the box, we considered possible differences in search duration across situations to be indicative of infants having encoded the other person's belief, and this in turn having modulated infants' search behavior. As it is unclear how the set-size limit of infants' object tracking system would interact with

the tracking involved in representing belief contents, we aimed at calculations where the overall number of object-files would not exceed the set-size limit of 3 items.

We tested whether 14-month-old infants can track in a scenario multiple objects from someone else's perspective and represent the person's belief accordingly. We presented infants with scenarios where objects were hidden to, and retrieved from, an opaque box; and subsequently gave an opportunity for infants to search in the box. Infants were assigned to one of two conditions, depending on whether at the end of the scenarios an object remained in the box, or whether all hidden objects were retrieved. Additionally, we varied within subjects in the test trials the other person's belief regarding the content of the box. Through measuring infants' manual search duration across these scenarios we aimed to measure whether infants successfully represented the other person's belief regarding the content of the box, and whether this manifested itself in an indirect measure, namely a modulation of infant's search durations by the other person's belief.

The present study was therefore motivated by three broad questions. First, whether infants' ToM capacities can deal with scenarios where they need to track another person's representation of more than one object. Second, whether this capacity is there from around the age when they were found with similar measures to show these abilities for their own perspective. Third, whether we can find an effect of another person's beliefs on infants' own behavior, manifested in an active behavioral paradigm.

3.2.1 Methods

Participants

All infants were recruited through a local database and parents signed an informed consent prior to participation. All studies received full ethical approval from the United Ethical Review Committee for Research in Psychology (EPKEB) in Hungary and were conducted according to the principles of the Declaration of Helsinki.

Sixty-four healthy full-term 14-month-old infants participated (32 per condition; age range from 14;0 [Months; Days] to 15;1, mean age = 14;17 in FB1 condition; and age range from 14;5 to 14;29, mean age = 14;17 in FB0 condition); 41 were girls (20 in FB1 and 21 in FB0). Twenty-six additional infants were tested but not included in the analyses because they did not search in any of the trials (14), the study was not completed because the baby fussed out (9), or due to experimental error in the procedure or error in the recording (3).

Materials

We used a white cardboard box (29*29*15 cm) with a 14*8 cm opening that was covered by an elastic cloth that prevented infants from seeing inside the box but enabled reaching into it. The box had a hidden compartment in its back where objects could be put in but not taken out. Objects could be taken out from the compartment through a back opening of the box that was not visible and was fastened throughout the experiment.

In Familiarization we used colorful whistles with a ball inside that made a rattling sound. In Test there were whistles of different colors (red, green, or blue) that did not rattle (see Figure 3.1).

Figure 3.1. Experimental objects

Procedure

Infants sat on their parent's lap, at an 80*60 table with two experimenters sitting on the two longer sides of the table (See Fig. 3.2). Two cameras recorded the experiment from the infant's left and right side.

Each session began with three Familiarization trials, followed by two Test trials. Test trials were either False Belief (FB) or True Belief (TB) trials. Each infant participated in one FB and one TB trial; the order of test trials was counterbalanced between participants.

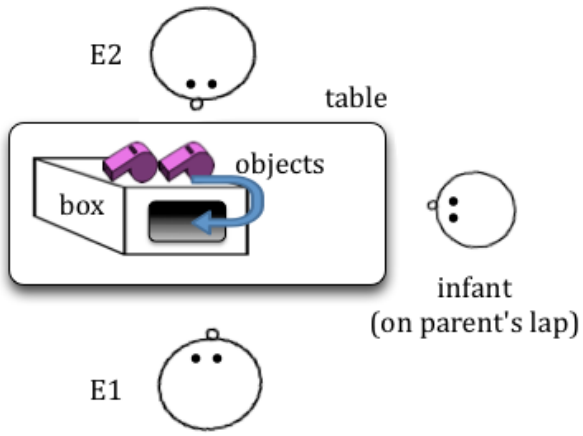

Figure 3.2. Schematic drawing of setup during Experiment 3.

Familiarization trials

In Familiarization only E1 interacted with the infant. She showed the box to the infant and said: "See? I have a box here. Look, it is a nice box. You can reach into it, like this. Do you want to try?". She let the infant reach in the box if he wanted to. She then said "Look, I'm going to show you something!", and took out a whistle from a small bag she wore on her other side than the infant was sitting. Then she put the whistle on the top of the box, pointed at it and said: "Look!", and she put it into the box. Following this, in the first trial she asked "What's in the box?", reached into it, and while she was moving her hand around, she said "I'm searching for it… searching for it". When she retrieved the whistle, she demonstrated how it rattled and blew the whistle, then handed it to the infant to briefly explore, and finally took it back and put it away in a small bag under her chair. In the second and third Familiarization trials after putting the whistle in the box, the experimenter encouraged the infant to search for it, by saying: "What's in the box? Will you search for it?". If the child was reluctant to reach into the box, the experimenter told the parent to retrieve the whistle. This served as enabling the child to reach into the box. Infants were encouraged to search

until they reached into the box at least once during Familiarization; successful retrieval of the whistle was not a prerequisite to proceed to Test trials, as this was influenced by a chance factor (whether the whistle happened to be closer to the infant's hand). If infants could not retrieve the whistle but searched for it, the experimenter retrieved it from the box and handed it to them.

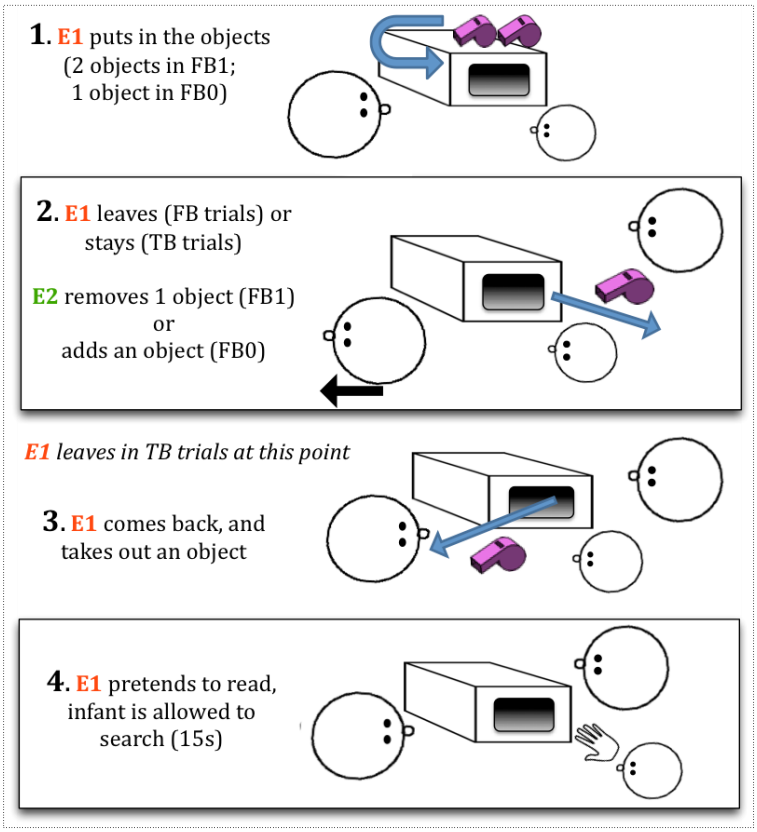

Figure 3.3. Schematic depiction of procedure in Experiment 3. Images show events in FB1 condition; for differences in FB0 condition see text in figure, and main text.

Test Trials: FB1 condition

False Belief Test trials. Test trials started as the Familiarization trials: E1 took out this time two (non-rattling) whistles from her bag, she put the whistles on the top of the box, pointed at them and said: "Look!", then she put them into the box (see Figure 3.3). Following this, in FB trials E1 reached to her pocket and said: "Oh, my phone is

ringing, I have to run out", and left the room. In her absence, E2 called the infant's attention saying "Look!", reached into the box, retrieved one of the whistles and said: "I am taking this one out". Following this, E1 came back, sat down and performed the same searching action as in the first Familiarization trial: retrieved the whistle, blew it and put it away. Finally, she took a book from next to the box and said: "I have to look up something now", pushed the box to the side to make room for her book, and as a result the box landed in front of the infant. Then she pretended to read for 15s. For this time E2 did not interact with the infant either. Infants were hence not encouraged at this point to search for an object but were merely given the opportunity, as a by-product of E1's other activity. After the 15s E1 looked up, put down the book and said: "Ok, we are done". If the FB trial was first, then to avoid "cumulating" beliefs, before continuing to the next Test trial E2 took out (from a shoulder bag she was wearing) the whistle she retrieved previously, gave it to E1 and said: "Look, I still had this". After this the second Test trial followed.

True Belief Test trials. TB trials were identical to FB trials with one difference: E1 stayed after putting the whistles in the box, but left after E2 retrieved the whistle, and came back right away. When she came back, she continued as in FB trials with retrieving one whistle. Crucially, while in FB she thought she is still leaving one behind in the box, in TB she knew that when she took out the whistle both whistles were taken out. If TB trial was first, before continuing to the second Test trial E2 took out the whistle she took, gave it to E1 and said: "Look, I still had this". This was not strictly necessary as E1 also knew she had the whistle with her, but it was matched to the FB-first trials in order to avoid effects due to this manipulation. After this the second Test trial followed.

Test Trials: FB0 condition

The FB0 condition was identical to the FB0 condition in the overall procedure, but differed in the number of whistles that were put in or taken out during the critical events in the test trials. Initially E1 always put in just one whistle in the box. In her absence (FB) or presence (TB) E2 put in another whistle while saying, "Look! I am putting this one in!", hence at this point there were two whistles in the box. When E1 came back she retrieved one of the whistles (which in FB she thought was the only one in the box). However, E2 actually put it in the secret compartment, therefore when in the end infants were allowed to search they believed a whistle to be in the box but they would never actually find it. This was matched to the procedure of Feigenson & Carey (2003) and was done so in order to avoid chance factor (when infants happen to find the toy) to determine when the search ended.

Coding

Search duration

We coded the duration of infants' searching during the 15s period when the box was in front of them. From infants' point of view, in the FB1 group at this point the box was expected to be empty, for the FB0 group it contained an object. Infants were scored as searching whenever one or both hands were in the box, with their knuckles past the entrance cloth of the box. If infants reached in to this extent but clearly just manipulated/played with the cloth we did not count it as searching. Forty percent of infants in each condition (FB1 and FB0) were coded by a second coder who was blind

to the purpose of the study; inter-rater agreement was $r(26)= .933$, $p<001$ for condition FB1, and $r(26)= .988$, $p<001$ for condition FB0.

Pointing and reaching

As an additional measure we also coded infants' pointing and reaching pattern added together from two time intervals: from the time point when E1 re-entered the room until she retrieved the object, which was approximately 10-15s and during the time when infants had the opportunity to search, which again was 15 s.

Through coding infants' pointing and reaching behavior we aimed to assess whether they communicated with the experimenter(s) in order to (i) ask for the object, or to (ii) communicate to E1 regarding her false belief. Behavior indicating (i) would include pointing or reaching towards the box, which should be more frequent in the FB0 condition compared to FB1 condition (as in the former there is indeed an object in the box to ask for). Conversely, there could be more pointing to E2 in FB1 condition TB trials than in FB0 condition in TB Trials (as in FB1 condition E2 has the last object). Behavior indicating (ii) would include pointing towards E2 in FB1 condition, or pointing towards the box in FB0 condition.

3.2.2. Results

Manual search

Infants' search duration was analyzed with regard to the other person's belief (whether one of the objects remained in the box) at the time when infants had a chance to search. In 50% of all the test trials, the other person believed that one of the objects was still in the box (hereinafter 'other believes =1', or OB1), in the other 50% she believed all objects have been retrieved from the box (hereinafter 'other believes =0', or OB0). In the FB1 condition 'other believes=1' referred to the false belief trials, and 'other believes =0' corresponded to trials when the other person knew the true state of affairs. Reversely, in the FB0 condition in 'other believes =1' trials the other person correctly represented that one of the objects remained in the box; whereas in the 'other believes =0' she had a false belief that there was only one object which was later taken out, and therefore the box did not contain an object at the moment of the infants' opportunity to search.

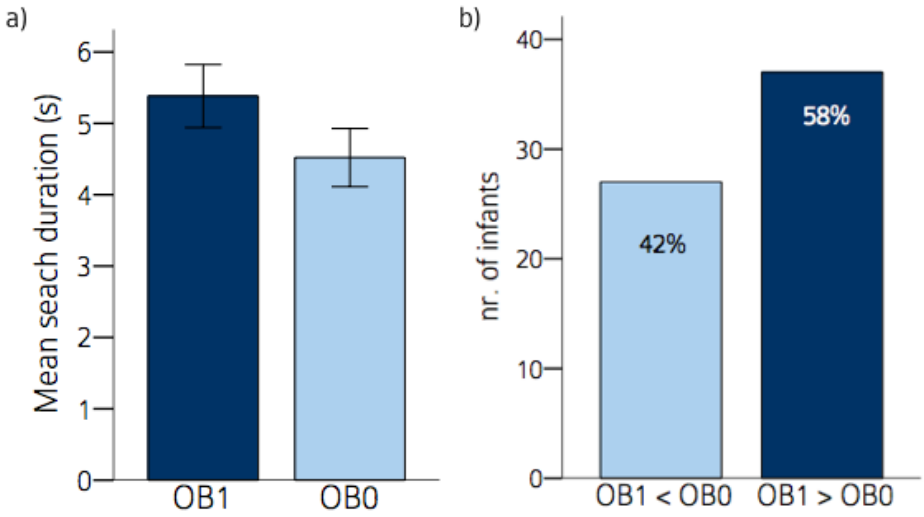

Figure 3.4. Infants' search behavior in Experiment 3. a) Search times in 'other believes=1' (OB1) and 'other believes=0' (OB0) trials. Error bars depict +/- 1 Standard Error. b) number and percentage of infants who searched longer in OB1 trials (OB1>OB0) vs. who searched longer on OB0 trials or equally long in the two trials (OB1<OB0).

In order to assess whether the belief of the other person modulated the duration of infants' search, we analyzed search times in a 3-way mixed ANOVA with Belief ('other believes =1' vs. 'other believes =0') as within-subjects factor and Condition (FB0 vs. FB1) and Order of Test (FB first vs. TB first) as between-subjects factors. There was a significant main effect of Belief ($F(1, 60)=4.067$, $p=.048$, partial $\eta^2=.063$); infants searched longer in 'other believes 1' trials ($M_{OB1}=5.382$ s, $SE=.429$) than in 'other believes 0' trials ($M_{OB0}= 4.52$, $SE= .397$). For mean search durations, see Figure 3.4.

There was also a significant Belief * Condition * Order of Test interaction ($F(1, 60)=0.629$, $p= .003$, partial $\eta^2=.138$). To resolve this interaction, we performed separate 2-way ANOVAs in both conditions, with Belief ('other believes =1' vs. 'other believes =0') as within-subjects factor and Order of Test (FB first vs. TB first) as between-subjects factor. In both conditions there was a significant interaction between Belief and Order of Test (in FB1 condition: $F(1, 30)=4.162$, $p= .05$, partial $\eta^2=.122$; and in FB0 condition: $F(1, 30)=5.537$, $p= .025$, partial $\eta^2=.156$). This interaction was due to the fact that in FB1 condition the effect of Belief was significant in the FB First order ($F(1, 15)=5.1$, $p= .039$, partial $\eta^2=.254$), whereas in the FB1 condition it was significant in the TB First order ($F(1, 15)=7.763$, $p= .014$, partial $\eta^2=.341$). There was also an overall effect of Order ($F(1, 60)=5.101$, $p= .028$, partial $\eta^2=.078$), with longer search durations in FB first than in TB first ($M_{FBfirst}=5.751$ s, $SE=.501$, and $M_{TBfirst}= 4.152$, $SE= .501$, respectively).

If analyzed by trials separately, the above interaction was mirrored in the trials in which infant's search times were modulated by the other's belief, depending on the

condition (see Figure 3.5). In the FB1 condition there was a significant difference in search duration on the first trial between infants who received OB1 or OB0 trials ($t(30)=2.302$, $p=.028$, Cohen's $d= 0.815$), with longer search times on OB1 trials compared to OB0 trials ($M_{OB1}= 6.9$ s, $SE= .843$; $M_{OB0}= 4.328$ s, $SE= .733$). The effect on the second trial was not significant. In the FB0 condition, the effect on the first trial was not significant, but there was a significant difference in search duration on the second trial between infants who received OB1 or OB0 trials ($t(30)=2.302$, $p=.028$, Cohen's $d= 0.768$), with longer search times on OB1 trials compared to OB0 trials ($M_{OB1}= 5.388$ s, $SE= .993$; $M_{OB0}= 3.037$ s, $SE= .433$). The effect on the first trial was not significant.

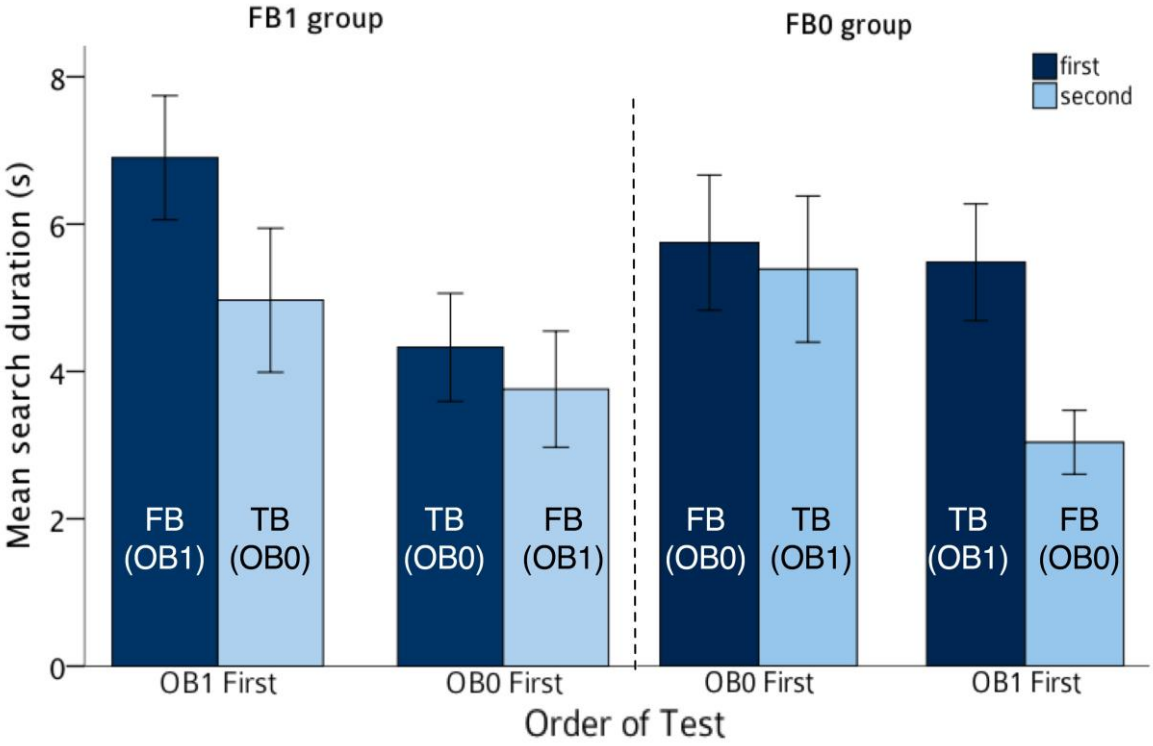

Figure 3.5. Infants' search behavior in Experiment 3. Search times in the first and second test trial, by Order of Test (OB1 First, or OB0 First), in the two Conditions (FB1 group and FB0 group). False Belief and True Belief trials are denoted as FB and TB; and 'Other believes 1' and 'Other believes 0' as OB1 and OB0, respectively. Error bars depict +/- 1 Standard Error.

When analyzed across all trial types, there was no effect of Condition (FB0 vs. FB1). These two conditions differed with regard to the actual number of objects in the

box at the time of infants' search: in FB0 there was in fact an object to search for (we will call this condition for the present analysis 'FB0/object remains'), whereas in FB1 the box was empty (we will call this condition for the present analysis 'FB1/box empty'). To further analyze whether this difference manifested itself in infants' search times, we analyzed the TB trials from each condition. While infants indeed seemed to search longer in the 'FB0/object remains' ($M_{FB0/object\_remains}$=5.435 s, *SE*=.625) than in the 'FB1/box empty' condition(M_{FB1/box_empty}=4.647 s, *SE*=.604), this difference was not significant. This difference manifested itself also when only those infants' values were included who received TB trial first, in order to avoid carryover effects from FB trials, but the difference remained non-significant ($t(30)$=-1.068; p=.294).

In sum, infants overall searched longer in test trials where the other person believed an object to be in the box (OB1 trials) compared to trials where she believed all objects have been removed from the box (OB0 trials). This effect manifested itself differently in the two groups corresponding to the real states of affairs (in the FB1 group there was no object in the box, whereas in the FB0 group there was an object inside). In the group where there was no toy in the box at the time of search, the effect of the others' belief appeared on the first trial (between infants). Conversely, in the group who did in fact have a toy in the box, the effect was significant on the second trial, between infants. Lastly, there was no main effect in search duration between FB1 (box empty) and FB0 (box contains object) groups.

Pointing and reaching

Infants' pointing and reaching behavior was coded in order to assess whether they communicated about the content of the box. Overall infants pointed/reached to either the box (53.1% of infants), or to E1 (29.7% of infants), and almost no infant pointed or reached towards E2 (9.4% of infants). Significantly more infants signaled to E1 compared to E2 (p= .015, binomial test), and significantly more infants signaled to the box compared to E2 (p<.0001, binomial test). There was a trend towards more infants signaling towards the box compared to E1 (p= .053). Signaling to E2 was therefore not analyzed further, due to the small number of instances of signaling to E2.

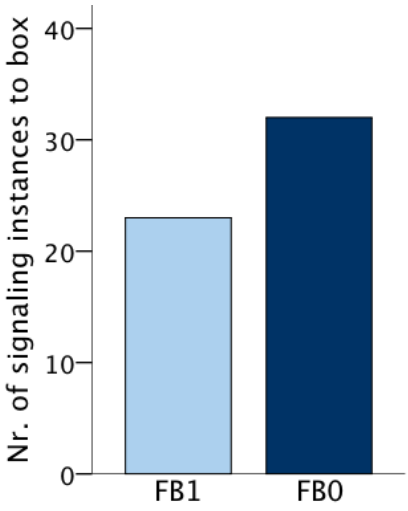

Figure 3.6. Total number of instances when infants signaled (pointed, or reached) towards the box in FB1 and FB0 conditions, in Experiment 3.

We analyzed signaling towards the box in order to see whether infants showed any difference between conditions (FB0 vs. FB1). Altogether, more infants signaled towards the box in FB0 condition (n_{box_FB0}=21) than in FB1 condition (n_{box_FB1}=13), and there were more overall signaling instances in FB0 condition (n_{sign_FB0}=32) than

in FB1 condition (n_{sign_FB1}=23), however, these differences were not significant (see Figure 3.6), and there were no other significant differences.

3.3. Discussion

The present study aimed to tap onto three broad questions. First, it asked whether infants' ToM could deal with scenarios where they need to track another person's representation of multiple objects. Second, it aimed to test whether this capacity is there from around the age when in similar paradigms infants demonstrated their abilities to track multiple objects from their own perspective. Finally, it investigated the above questions through a potential modulation effect of another person's beliefs on infants' own behavior in an active behavioral paradigm.

We presented 14-month-old infants with scenarios where they had to track the location of two objects that were hidden in an opaque box and could be retrieved sequentially; and measured infants' manual search durations at the end of various scenarios. We manipulated the access of another person to the events, and as a result in one event within a scenario pair the other person had a true belief about the content of the box, and in the other case she had a false belief about what is inside.

To assess whether infants represented the other person's belief regarding the content of the box we compared search durations within scenario pairs: in trials when the other person believed an object to be present in the box, and trials where she believed the box to be empty. We reasoned that if infants represent the other person's belief, then this attributed representation might modulate their own behavior indirectly. This would manifest itself in longer search durations in trials

when the other person believes an object to be present. Consistent with these predictions, infants indeed showed longer search duration when the other person believed an object to be in the box, compared to when she believed it to be empty. Since the only difference between these scenarios was what the other person believed to be the case, we take this as evidence that infants represented the other person's belief about the content of the box.

The fact infants' own search was modulated by what someone else believed to be true is in itself quite telling of the way beliefs are represented in the infant mind, as it supports the notion that own and attributed representations are highly similar in format. By highly similar we mean that they (i) respond to the same kinds of inputs (i.e. equivalent changes in the environment relevant to the person to which a certain representation 'belongs'), (ii) might be at least partially handled by the same cognitive systems within the brain (which is in line with results from Chapter 2), and (iii) feed into the same subsequent processes (prediction of others' actions, and infants' own action planning).

Additionally, between conditions the true state of affairs differed; in one group there was indeed an object in the box at the end of the scenario (FB0 group), in the other there was not (FB1 group). Therefore overall longer search durations in the FB0 compared to the FB1 group (where all objects have been retrieved) would signal that infants' search duration is driven by between-group differences in representing the true state of affairs. However, the difference in search times between groups (FB0 and FB1) was not significant, which suggests that between-groups differences did not grasp whether or not there was indeed an object present in the box.

Infants' pointing and reaching frequency gives tentative evidence suggesting that the two groups (FB0 vs. FB1) differed in their inferences about the box. Infants in the FB0 group, where there was indeed an object in the box, tended to signal more towards the box, which could possibly be interpreted as requests for the object, or informing the experimenter about the presence of the object in the box. Contrary to other studies showing spontaneous helping of others who have a false belief (Knudsen & Liszkowski, 2012b), in the present study there was no sign of helping or warning the experimenter (which would have been evident in the two groups signaling selectively towards the experimenter who had a false belief). It should be noted however, that this study was not set out to assess pointing behavior (the arrangement of the box, of the experimenters, and the infant was not suitable to measure selective pointing towards one or the other), therefore this additional analysis should only be taken at most as preliminary data that could be further disentangled with an appropriate design.

There was an additional difference in the data pattern between the two groups of infants (FB0 and FB1). In the FB1 group, when the box was empty but in false belief trials the other believed an object to be there, the difference between trial type was observed in the first trial (if analyzed between subjects); whereas in the FB0 group, where there was an object in the box but in false belief trials the other believed it to be empty, the effect was prominent on the second trial (between subjects). This difference was not predicted, but nevertheless can be explained with a different pattern of possible inferences and motivational factors in the two conditions. In the FB1 group (where there was no object in the box), a modulation effect would consist of *longer* search duration in false belief trials (when the other believed that there is one left), than in true belief trials (where the other person knew that all objects were

retrieved). Since infants might be more motivated to search in the beginning of the experiment, the modulation effect might be more prominent on the first trial. However, on the second test trial, they have already once failed to find the object (as during test trials they never find anything; even in cases when there should be an object, the object is in fact hidden in a hidden compartment in the back of the box). Therefore, on the second test trial their motivation to search might decrease to an equally low level on both kinds of trials, and the effect of the other's belief might not be observed. On the other hand, in the FB0 group (where there was indeed an object in the box), because their motivation to search is initially high; infants might search equally long on the first trial, regardless of the other's belief; but will be more easily pushed towards searching *less* in the second trial in false belief trials (when the other person believes there are no more objects left), when their motivation drops after the absence of success on the first trial, and gets distanced from a possible ceiling allowing for modulations to take place. However, such potential factors need to be clarified in further studies. To our knowledge there is no mention in other studies using the manual search task, whether infants would show less proclivity to search as the trials progress (Feigenson & Carey, 2003; Feigenson & Halberda, 2004; Zosh & Feigenson, 2012). Nevertheless, there is a further difference between the previous studies and our paradigm; the above studies allow infants to find the first object(s), and this successful search is followed by the test phase(s). In the present study, it was always the experimenter who retrieved or added the object before the search phase. This, while a subtle difference, may contribute to infants' involvement in the process.

How did infants represent the other's belief content in our scenarios? In descriptive terms, in the FB1 group infants likely represented something like 'the other person believes that there is an object in the box'. This representation of the

object in the box made infants search longer in these trials. It is less clear, what infants represented in the false belief trials of the FB0 condition, when the other person thought all objects had been retrieved. It is a matter of debate what infants represent from their own point of view when all objects that have been hidden in front of them, have also been retrieved (and therefore the box is empty). Infants could (i) represent the box to be empty (as a feature attached to the box's representation; cf. Mody & Carey, 2016), (ii) represent the objects to be elsewhere (e.g. through their object-tracking system), (iii) use negation to represent that the objects in question are *not* in the box (Austin, Theakston, Lieven, & Tomasello, 2014; Nordmeyer & Frank, 2014), (iv) represent it as an empty set, i.e. the numerical quantity zero (while to date the earlies evidence in humans comes from preschool years, see Merritt & Brannon, 2011; there is evidence that monkeys' parietal cortex can represent numerosity zero: Okuyama, Kuki, & Mushiake, 2015), or simply (v) not have represented anything to be the box. Similarly, when they attribute to the other person that all the objects she has seen to be hidden were also retrieved, in theory they might attribute to her any of the above. However, if they would have simply not attributed any belief to the other person, that could not make them search less in these cases, as it is hard to imagine how the absence of a representation could have an effect on behavior. Therefore given that they searched *less* on the false belief trials, this would involve the attribution of a representation of some sort; such as negation of presence of the object, or representing emptiness.

Alternatively, in the FB0 group it could be that having two positive beliefs[15] about the object in the box in true belief trials (one for themselves and one for the experimenter) might prompt them to search longer; compared to how much they would search by default 'just' due their own belief in false belief trials. Since there were no baseline trials (without any other agent present, if only the infant believed an object to be in the box, or believe it to be empty), we cannot tell whether one should think about this modulation as infants searching *less* when someone else has an additional 'negative belief' (cf. Footnote 15), or search *more* when someone has an additional 'positive belief', compared to when it is only their own reality representations at play.

What representations do infants build on when tracking beliefs in these scenarios? It is possible that they might use representations described as 'registrations' (see Introduction of the present thesis, Chapter 1.1.2.), namely relations that simply track the objects themselves (Butterfill & Apperly, 2013; Rakoczy et al., 2015). However, it is unclear whether representing registrations could explain the modulation effect. For it is doubtful that registrations could be 'mistaken' for one's own representations, as they are different from infants' reality representations. Therefore unless one supposes that infants represent the environment form their own perspective with registrations; the modulation effect is incompatible with registrations subserving tracking mental states, and suggests that infants metarepresent the other person's belief content.

[15] By positive belief we mean the belief that there is an object in the box (compared to the negative belief that the object is not there, or that there is no object in the box)

As from their own perspective infants were argued to track objects via object files, these findings raise the question whether belief contents are subject to similar limitations as object-files are. Infants in this study had to track someone else's belief regarding up to two objects. Previous studies confirmed that from their own perspective infants can track 3 items, but their performance breaks down at 4 (Feigenson & Carey, 2003, 2005). Based on such findings it was argued that infants in these situations track objects via object-files that are indexed to the objects, and show a signature limit of 3 indexes that can be tracked in parallel. In the introduction of this thesis a possible mechanism was described, where infants would track others' representations via mental files ascribed to another person; and such mental files would be connected to infants' own object-files. In theory, this would enable tracking every object the infant tracks from their own perspective, also from another person's perspective. The object-file system's limitations were suggested to arise specifically from features of the visual system (i.e. on the limits of visual attention); therefore in principle the mental-file representations are not subject to such limitations. Hence the nature of mental files would in theory allow for tracking any number of agents' beliefs about an object. However, memory limitations would likely not allow tracking beliefs above a certain number. In fact, it has been suggested that restrictions on the number of object-files and items in working memory are due to the same underlying capacity limit (Cowan, 2001). Convergent findings come from studies building on the phenomenon of chunking in working memory; i.e. that binding individuals into sets enables to overcome working memory limitations (Miller, 1956). Specifically, using the manual search paradigm Feigenson and Halberda (2004) found that infants can overcome their limitations of tracking only 3 items (a limitation that was suggested to be a characteristic of indexing in the visual system), if the items could be grouped into two separate sets. In line with this, recent findings suggest that infants can use social

information (agent's affiliation) to chunk items in working memory, again allowing them to overcome the limit of 3 (Stahl & Feigenson, 2014).

These findings have two consequences. On one hand, if the object-based attention and working memory bounds come from the same underlying capacity limitation, then these limits will likely affect infants' attributed belief representations as well. On the other hand, chunking may take place through grouping representations into one's own versus others' representations (for example, regular mental files vs. vicarious files). This could allow overcoming the capacity limit in scenarios where one's own and the attributed representations sum up to more than 3 items. How working memory, object tracking and mechanisms of belief representation interact with each other, needs to be therefore subject of future research.

In sum, the present chapter has demonstrated that 14-month-old infants can track multiple objects as the content of someone else's belief; and can perform simple arithmetic operations on these belief contents. Infants showed indication of this capacity at the same age when in a similar setting they demonstrated such abilities from their own perspective. In addition, this manifested itself in a modulation effect, whereby infants' own search times for a potential hidden object were influenced by another person's belief about the presence or absence of the object. Together, these findings provide supporting evidence that infants build on their cognitive apparatus that enables tracking objects in the environment, to also represent others' beliefs about these objects. Moreover, infants showed the modulation effect in an active behavioral paradigm, which corroborates previous findings that have found the modulation effect in young infants using looking-time measures (Kovács et al., 2010).

The present study tested infants' capacity to infer that another person individuates and tracks *two* objects based on *spatiotemporal* information. Further studies may investigate how infants can track three (and possibly more) objects, as well as the richness of the object representations infants can attribute to others, and the various kinds of information that can be the basis of such inferences. Chapter 4 of this thesis presents two studies that investigate infants' abilities to attribute to others beliefs based on individuation through feature/kind information.

Chapter 4

Attributing beliefs involving individuation of objects based on feature/kind properties

4.1. Introduction

When we track objects around us, the mind has a complex task to solve: it has to identify which object is being tracked while it is in sight, and at every occlusion and re-emergence of the object it has to decide whether it is the same object as what was seen before. Similarly, in order to successfully represent someone else's mental states about an object, one has to (i) individuate and at each additional encounter re-identify the object (while tracking it through occlusion), and (ii) assume that the other person also performs these operations. Any location-change false belief task, for instance, has to involve the attribution of individuating one object at, say, location A; and the (possible) re-identification at a later time point. Even when more objects are involved, in the majority of cases this is possible through tracking spatiotemporal properties, such as witnessing the two objects in two different spatial locations at the same time. Such tracking also enables success in Chapter 3: the objects are presented in a way that enables individuation based on spatial location both from the other person's perspective, and from the infants' point of view. However, spatiotemporal information is not always available, therefore being able to rely on other types of information for individuation in social contexts may be of essence.

In the present chapter we explored cases where tracking spatiotemporal information is not enough to successfully predict individuation from someone else's point of view. We aimed to investigate whether infants can make correct inferences about someone else's belief, when this is based on the assumption that they

(i) not only can track each individual object based on spatiotemporal information (as shown in Chapter 3),

but also that they

(ii) track features of these objects, e.g. their appearance; and

(iii) individuate the objects based on their features.

Experiment 4 explores whether infants can successfully infer that someone can have a *false* belief based on a *correct* individuation of two objects that are of different appearance. Building on this, Experiment 5 asks whether infants can ascribe to someone a false belief based on that person's *mistaken* individuation of two appearances of one object into two objects. The latter question, namely whether infants can ascribe to others false beliefs about object identity, has been of great interest to researchers with regard to mental state understanding.

Individuation based on feature/identity information from first-person perspective

In everyday life we constantly need to monitor objects in our environment. Imagine that you are tidying up and you find two identical pens on your desk, which you put into your drawer. Later you need to write down something, and take out one of these pens from the drawer. If you correctly tracked the spatiotemporal information, you would infer that there must be two objects, and you would know that the other pen is still in the drawer. Therefore if you needed it, you would search for it in there. As shown in the previous chapter, infants are not only able to make such inferences themselves, but also successfully attribute them to others at least by the age of 14 months.

Now imagine a scenario where you put your pen in a drawer, and close the drawer. Then your phone rings and you have to quickly go out of the room to take the

call. When you come back, you open the drawer again, reach into it, and find and take out a straw. What inference would you make about this drawer? Most likely you would think that the straw got in there somehow (or was there before), and that your pen is still in there. This inference would not be based on distinctive spatiotemporal information (as at any given time you only saw one object). It would be based on the fact that the straw *looks* different; that is, it has a different featural property (its appearance/function) as the pen does. Hence you will reason that there have to be two different objects, one that you put in the drawer (the pen), and one that you then retrieved (the straw). If you had the chance to open the drawer again, you would open it and search for your pen. However, what you don't know is that while you went out and took that call, your friend decided to play a trick on you, and exchanged your pen for a straw. Therefore you will end up with a false belief that your pen is still inside that drawer, when in fact the drawer would be empty.

For the observers, in order to correctly infer what you think about the content of the drawer, they would need to understand these inferences you likely made. From their own point of view if they would track the 'movements' of these two objects - the pen and the straw –, then successful tracking would result in the final inference that the drawer is empty. To represent your belief about the drawer, they would however need to understand that you (correctly) think there are two objects involved, but also understand that the lack of knowledge about the exchange will lead you to (mistakenly) believe that one of them is still inside. It would lead you to this conclusion because you not only track the number of objects that you saw go in versus come out from the drawer; but that you likely keep track that the object that you put in is *another* object than the one you later took out – therefore the most likely case is that the first one is still somewhere inside the drawer.

Infants are able to individuate objects not just based on spatiotemporal information but based on feature/kind information as well, already at 12 months of age as shown by their looking pattern (Xu & Carey, 1996), and at 18 months when tested with the manual search paradigm (Zosh & Feigenson, 2012). Zosh and Feigenson (2012) found that 18-month-old infants search less when they are shown that for instance two objects are put in a box, and then two objects of the same appearance are retrieved (no-switch trials); compared to when one of the objects retrieved is different in appearance (switch trials). Differences in features were indicative of the objects being from two different kinds (e.g. a cat and a car). Infants' search behavior suggests that in the switch trials they successfully individuated altogether 3 objects (presumably of different kinds), and inferred that after the retrieval of two, one must have remained in the box.

Representing beliefs involving individuation based on feature/identity information

The present studies aimed to extend findings from Chapter 3, and investigate whether infants can use individuation based on feature/identity information to represent others' beliefs. While Zosh and Feigenson (2012) tested 18-month-olds, we aimed to match the age of infants to those in Chapter 3. Earlier, 12-month-olds were found to individuate objects based on kind-relevant information (Xu & Carey, 1996), and 14-month-olds search behavior suggests successful tracking of up to 3 objects (Feigenson & Carey, 2003). Therefore we reasoned that at 14 months of age we can

expect them to show search behavior in line with individuation based on kind information as well as the attribution of such representations to other people.

4.2. Experiment 4: Attributing false beliefs based on correct individuation by features of objects

In Experiment 1 we asked whether infants can successfully track feature/kind information of multiple objects, and attribute to others the individuation and tracking of these features as well as subsequent inferences based on such individuation. We created scenarios where an object was placed inside a box – visibly to both the infant and another person. Following this, the infant always observed that this object was exchanged for another one, but we varied across trials whether the other person sees this exchange. Finally, the second object was taken out (visibly to both the infant and the other person), leaving the box empty.

We reasoned that if infants (i) track features of the object, (ii) assume that the other person tracks these too, (iii) track whether the other person sees the exchange, and (iv) infer whether the other person has individuated one or two objects in the box, then they should ascribe different beliefs to the other person in the two trials. If the other person has not seen the switch (unseen-switch trials), infants should infer in the end that the person mistakenly thinks there is still an object in the box; whereas if she was aware of the switch (seen-switch trials), she should now know that the box is empty.

Our dependent measure, like in Chapter 3, was infants' search duration in the two kinds of trials. Since we found a modulation of infants' own search behavior depending on the other's belief (whether she thinks there is an object in the box or not), if in the present case infants accurately represent the other person's belief based on her experience, we predicted that they would show a similar modulation effect. In the unseen-switch test trials the other person believed that one of the objects was still in the box (hereinafter we call these trials 'other believes =1', or OB-1); whereas in the seen-switch trials the other person knew that there is no object in the box (hereinafter 'other believes =0', or OB0). Hence we reasoned that longer search duration in the unseen-switch (OB-1) trials compared to the seen-switch (OB0) trials would indicate that infants are sensitive to the social partner's beliefs in these scenarios.

Additionally, we were interested in whether we would replicate the interaction between the modulation effect and test trial to that of the FB1 group in Chapter 3. In Chapter 3, as well as in the present study a potential modulation effect would entail *longer* search duration in false belief trials (as there other person believes an object to be in the box, while in fact the box is empty). Since infants might be more motivated to search in the beginning of the experiment, this longer search duration might be more prominent on the first trial. However, on the second test trial, they have already once failed to find the object (as during test trials they never find anything; even in cases when there should be an object, the object is in fact hidden in a hidden compartment in the back of the box). Therefore, on the second test trial their motivation to search might decrease and the effect of the other's belief might not be observed. Based on these considerations, we hypothesized that also in the present

study the modulation effect may be significant on the first trial between subjects, but not manifest itself on the second test trial.

4.2.1. Methods

Participants

Thirty-two 14-month-old infants participated (age range from 14;0 [Months; Days] to 14;29 mean age = 14;16); 12 were girls. 11 additional infants were tested but not included in the analyses because they did not search in any of the trials (8), the study was not completed because the baby fussed out (1), parental interference (1), or due to experimental error in the procedure (1).

Materials and Procedure, Coding

The hiding box, the whistles used in Familiarization, the setting, Familiarization Trials, and coding procedure were identical to those in Chapter 3. In Test, we used the same whistles as in Chapter 3, as well as little toy rattles to which the whistles were exchanged (See Figure 4.1).

Test Trials

In Test, E1 always put in just one whistle in the box. In her absence (FB) or presence (TB) E2 exchanged this whistle to a rattle while saying: "Look! I am exchanging this!". When E1 came back she retrieved the rattle, said: "Oh, how nice!", played with a little, and then put it away. Then, similarly to Chapter 3, infants had the

opportunity to search. All other steps of the procedure were identical to those in Chapter 3.

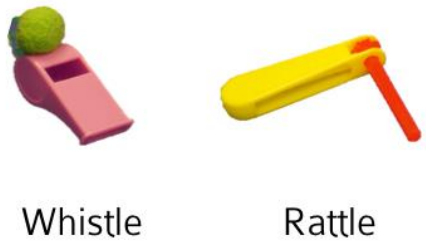

Figure 4.1. Objects used in Experiment 4.

Note that while the two experimental objects (whistle and rattle) were indeed of different kinds, this might not be evident for a 14-month-old infant, as both are relatively unfamiliar objects. In this Experiment we did not name the objects, therefore label information could not help infants in setting up two kind categories. However, we provided additional information to help individuation. First, in the familiarization trials whistles were used, albeit of slightly different appearance (see Figure 3.1 in Chapter 3), which provided infants experience with the function of the whistles (and some infants had experience from before, according to parental report). In addition, from the infant's point of view, spatiotemporal information supported individuating the objects: when the exchange of objects was performed, the experimenter retrieved the whistle, put it on the table, then took out the rattle from a bag, put it in the box, and then took the whistle and put it away. Therefore during a period of time, infants saw both objects at the same time. In seen-switch trials the other person had the same spatiotemporal information. In the unseen-switch trials the other person did not receive this spatiotemporal information, however, she was likely familiar with whistles (as she demonstrated their function in the familiarization trials – albeit different exemplars, but highly similar in appearance). Additionally,

when she reached into the box and retrieved the rattle, while uttering “oh, how nice!” she used the rattle to play some sound. This suggested that she was familiar with the function of the rattle. As from first-person perspective function information helps individuation in infants from early on (Futó et al., 2010), if they track that the other person has demonstrated different functions on the two objects, they can use this function information for individuation from her perspective.

4.2.2. Results

We entered search durations to a 2-way mixed ANOVA with Belief (OB-1 vs. OB0) as within-subjects factor and Order of Test (OB0 first vs. OB-1 first) as between-subjects factors. There was a marginally significant main effect of Belief ($F(1, 30)=3.684$, $p=.064$, partial $\eta^2=.109$; see Figure 4.2a); infants searched longer in ‘other believes 1’ trials ($M_{OB-1}=5.876$ s, $SE=.569$) than in ‘other believes 0’ trials ($M_{OB0}=4.807$, $SE=.626$). Out of 32 infants 21 showed the predicted pattern (see Figure 4.2b).

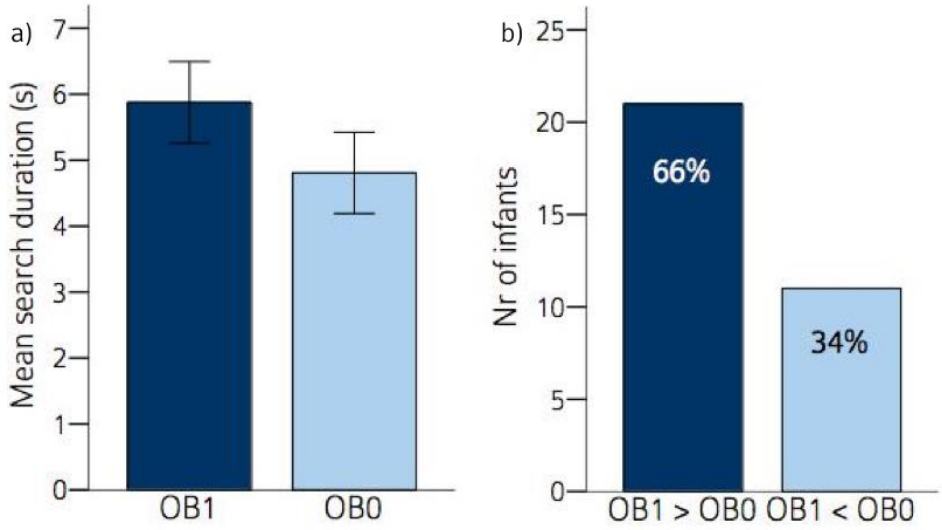

Figure 4.2. Infants’ search behavior in Experiment 4. a) Search times in 'other believes=1' (OB1) and 'other believes=0' (OB0) trials. Error bars depict +/- 1 Standard Error. b) number and percentage of infants who searched longer in OB1 trials (OB1>OB0) vs. who searched longer on OB0 trials or equally long in the two trials (OB1<OB0).

There was a significant interaction between Belief and Order of Test ($F(1,30)=7.026$, $p= .013$, partial $\eta^2=.190$; See Figure 4.3). In order to resolve this interaction, we performed separate one-way ANOVAs in both Order of Test conditions (OB0 first and OB-1 first), with Belief (OB-1 vs. OB0) as within-subjects factor. In the OB-1 first group there was a significant difference between OB-1 and OB0 trials ($F(1, 15)=8.364$, $p= .011$, partial $\eta^2=.358$), with longer search durations in OB-1 trials ($M_{OB\text{-}1}=7.346$ s, $SE=.665$) than in OB0 trials ($M_{OB0}= 4.801$, $SE= .984$). In the OB0 first group there was no significant effect of Belief ($F(1, 15)=.356$, $p= .56$, partial $\eta^2=.023$).

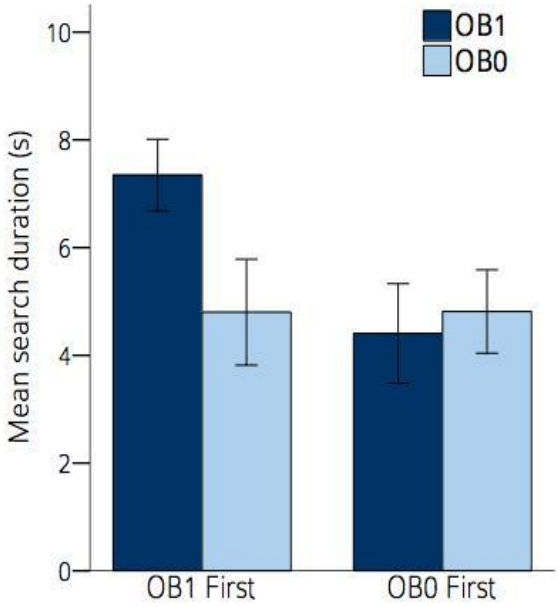

Figure 4.3. Infants' search behavior in Experiment 4. Mean search duration in 'other believes=1' (OB-1) and 'other believes=0' (OB0) trials, in the two Orders of Test trials. Error bars depict +/- 1 Standard Error.

If analyzed by trials separately this difference between the two groups (OB0 first and OB-1 first) manifested itself in between-group differences on the first trial, but not the second trial. On the first trial, there was a significant difference between infants who received the OB-1 trial and those who were presented with the OB0 trial first ($t(30)=2.482$, $p=.019$, Cohen's $d= 0.878$). This difference was not significant between groups on the second trial ($t(30)=0.292$, $p=.772$).

4.2.3. Discussion

In the present study we asked whether 14-month-old infants can successfully track feature/kind information of multiple objects; and assume that others encode this information as well as use it to individuate and track the objects.

Building on the methodology used in Chapter 3, we created scenarios where a person placed an object in an opaque box, and shortly after this, following a brief search in the box she took out a different object. Crucially, between this hiding and retrieval, the first object was exchanged by another experimenter for a second object, which event was either seen or not seen by the first person; but always observed by the infant. In the end the second object was taken out (visibly to both the infant and the other person), leaving the box empty. At this point infants had the opportunity to search in the box. We asked whether infants understand that if the person has not seen the switch (unseen-switch trials) then in the end she will mistakenly think there is still an object in the box (namely, the original object that she has placed there); whereas if she was aware of the switch (seen-switch trials), she should know that the box is now empty.

We measured infants' search durations in the unseen-switch and seen-switch trials. As in both of these trials infants themselves could know that the box was empty when they were allowed to search, we reasoned that any difference in search duration would reflect that infants are sensitive to the other person's belief about the content of the box, which had modulated infants' own behavior, as in the study of Chapter 3. Thus we predicted that if infants track the other person's belief about the content of the box and this has an effect on infants' behavior, this should manifest itself in longer search durations in unseen-switch (OB-1) trials, compared to seen-switch (OB0) trials.

In addition, we were interested whether the interaction between the effect of belief, and test trial that we observed in the FB1 group in Chapter 3, would replicate. There, on the first trial those infants searched longer who participated in a false belief (OB-1) trial, compared to infants who received the true belief (OB0) trial, but this between-group effect was not significant on the second trial. The interplay between potential motivational factors and the modulation effect should manifest itself analogously in the present study, as in both cases the modulation effect would entail *longer* search duration in false belief trials. Since infants might be more motivated to search in the beginning of the experiment, this longer search duration might be more prominent on the first trial. In the present study we therefore predicted a similar pattern. Specifically, we predicted a potential interaction between the other's belief and the experimental trial, with the other's belief's effect being more pronounced on the first test trial (between subjects), than on the second test trial.

Indeed, infants overall showed a tendency to search longer in the box on trials when the other person had reasons to believe that there is still an object in the box, compared to when she likely believed it to be empty. This effect was significant on the first trial between subjects (but not on the second), which corresponds to the data pattern obtained in the FB1 group in Chapter 3. Various motivational and attentional factors may contribute to this pattern, some of which were outlined in Chapter 3. Future studies may disentangle the potential contributing factors, for example by involving infants more throughout the process, and therefore keeping their willingness to search at a potentially more even level.

We are not aware of any low-level factors that may explain the findings in our studies. For instance, it has been argued that in false belief scenarios the re-appearance of the agent causes retroactive interference, which makes them forget the critical events (Heyes, 2014a). However, in the present study (and throughout in Experiments 4-6) the actor leaves in both test trials; and comes back right before the last event happens. Therefore if such an event disrupts infants' memory of the events, it should do so similarly in the two conditions. In addition, it is an open question whether infants perceive the context in the search phase as the experimenter prompting them to search. However, this is unlikely, as the experimenter pushes the box in front of the infant with the apparent goal to make space for herself to read; and she does not pay attention to the child during this phase. Moreover, this by itself would not explain the effects found in this study (or in Experiment 3), as the experimenter's behavior was identical in all test trials.

The present study provides converging evidence for the modulation effect of the other's belief on one's own actions, in an active behavioral paradigm in infants. In

addition, it gives further support to the claim that infants can handle others' belief representations with a variety of representational contents. This shows a relative sophistication of infants' ToM abilities. However, it should be noted that these findings, together with those of Chapter 3, are compatible with some theories positing simpler mechanisms to subserve mindreading in infancy, which were discussed in the introduction (Chapter 1.2.2.). Specifically, the minimal-ToM theory of Apperly and Butterfill (2013) posit representations called 'registrations', that represent relations between an agent and an object. Such representations enable tracking the location (and potentially some features) of the objects; therefore spontaneous location-change tasks may be explained through positing such representations in the infant mind. Similarly, our events in Experiments 3 & 4 may be potentially tracked by such representations.

However, registrations could not represent false beliefs about object identity. As they are relations between an agent and a physical object; they are incapable of handling aspectuality, i.e. that someone represent an external referent under some aspect (e.g. Superman) but not under another one (e.g. Clark Kent). Experiment 5 aimed to probe with the paradigm used in Experiments 3 & 4, whether 14-month-old infants can represent another person's false belief about object identity.

4.3. Experiment 5 – Attributing false beliefs based on incorrect individuation by features of objects

Experiments 3 & 4 in the present thesis provide evidence supportive of infants' abilities false beliefs based on correct individuation of multiple objects based on

spatiotemporal (Experiment 3) and feature/kind (Experiment 4) properties. This raises the question whether infants can attribute to others false beliefs based on *incorrect* individuation of objects. Incorrect individuation can take two forms: either as representing two objects mistakenly as one single object; or representing one objects incorrectly as two separate ones. Such mistakes can happen if there is no spatiotemporal information available, and one has to rely on other cues, such as appearance or other feature information.

Most objects have one stable appearance, therefore remembering its shape or color, or the category one takes it to belong to; can be used as a reliable basis for individuation. However, occasionally this can mislead one to individuate the wrong number of objects. Cacchione and colleagues (Cacchione et al., 2013) investigated 14-month-old infants' inferences regarding dual-identity objects – that is, objects that can change their appearance and therefore may seem to be two different objects. Infants observed object A being put in an opaque box; and object B retrieved. In fact, A and B were two forms of the same object, but observing this event should result in the individuation of two different objects, as shown by Zosh & Feigenson (2012). Indeed, infants who received a familiarization with regular objects (similarly to other manual search tasks) searched longer in such "switch" trials, than in an otherwise matched trial where object A was put in, and also the same form A was taken out (no-switch trials); which suggested that infants in switch trials mistakenly individuated two objects based on the two different appearances. However, there was another group of infants, who were initially familiarized with transformable objects (but not the same ones later used in test). This group searched equally on switch and no-switch trials, suggesting that in switch trials they likely inferred that the two appearances may belong to the same object, and therefore no object has remained to

search for. Together, these data confirm that 14-month-old infants can make two kinds of inferences. The findings from the group who was familiarized with regular objects ('no pretraining' group) confirm previous studies showing that infants at this age can use feature/kind change as the basis of object individuation (Zosh & Feigenson, 2012). The behavior of the group who was familiarized with the transformable objects ('pretraining' group) showed that based on previous knowledge, infants can make the inference that feature/kind change may not always be diagnostic for multiple objects. Note that the 'no pretraining' group had in fact a false belief about the identity of the object (i.e. that two 'senses' belong to the same 'referent'; cf. Chapter 1.1.2 of the present thesis, Perner's account); whereas the 'pretraining group' had information that justified the two appearances, and therefore had a true belief about the object. This raises the question whether infants use these inferences that two group of infants in Cacchione *et al.* (2013) likely made, to represent others' beliefs about object identity.

Tracking beliefs about identity of objects

Attributing beliefs about identity involves understanding the aspectuality of beliefs, and as such, it was suggested to be a key limitation of the early emerging, simple mindreading system responsible for spontaneous mentalizing (Apperly & Butterfill, 2009; Butterfill & Apperly, 2013). However, to date there is no satisfactory evidence that infants are incapable of such inferences, and few that would suggest that they do not happen spontaneously in adults. For example, Low & Watts (2013) found that kindergarteners and adults did not spontaneously represent an agent's false belief about the identity of an object, but they provided correct verbal answers

reflecting on these beliefs. However, as Carruthers (2015) pointed out, their task requires extensive executive control and working memory resources in order to perform several mental rotations during observing the scene; which explains well why participants nevertheless succeeded after some deliberation. Such task demands also apply to other tasks using similar scenarios (Low, Drummond, Walmsley, & Wang, 2014). This raises the possibility that if the external task demands are reasonably lowered, even younger infants may be able to deal with such tasks.

In fact, several findings suggest that at least by 18 months of age infants may successfully track similar scenarios. One such finding comes from Scott and colleagues (Scott & Baillargeon, 2009), who presented 18-month-olds with an agent observing two identically looking toy penguins. One penguin could be separated into two pieces, whereas the other couldn't. The agent always placed her key into the separable penguin, which she always encountered in its separated form, while the inseparable penguin was present in the conjoint form. In test the agent again was about to hide her key; this time the penguin that was visible was the separable penguin but this time in conjoint form, which infants were aware of (but the agent was not). Additionally, there was an object covered with a piece of clothing. An individual, who can represent false beliefs about identity, would likely have attributed to the agent that she believed the visible penguin to be the non-separable one. Therefore the agent should most likely reach for the object that is under the cloth, under the impression that that may be the separable penguin into which she wants to hide her key. Indeed; infants were surprised (as suggested by longer looking times during such events) if the agent reached for the conjoint penguin, suggesting that they successfully attributed to her the false belief that the conjoint penguin was the inseparable penguin. While Butterfill and Apperly (Butterfill & Apperly, 2013)

argue that these results can be explained by the infant attributing to the agent beliefs about object *kinds* (e.g. 'the agent assumes that a separable penguin is always present'); as Carruthers (2015) argues, this would involve ascribing quantified propositions (such as our example earlier in this sentence) to others, which again should not be possible with the simpler mindreading system.

Similar results were obtained with 14.5-month-olds in the study of Song & Baillargeon (2008) where infants were first familiarized infants with an agent who demonstrated a preference for one of object (a doll with blue hair) over another (a skunk). Then, unknowingly to the agent, the doll was put in a plain box, and the skunk was put in a box that had a tuft of hair on it, which was similar to the doll's hair. They reasoned that if infants can ascribe to the agent the false perception that the tuft of hair is the doll's hair, then they should expect her to reach for the box with the tuft. Indeed, infants looked longer to events when the agent correctly reached for the plain box, suggesting that it violated their expectations. While these studies suggest that infants were able to reason about another person's belief about the identity of an object; as Buttelman *et al.* (Buttelmann et al., 2015) point out, the above cases involve representing one representation for each object and confusing these two; whereas understanding aspectuality would involve maintaining two representations of one single object.

To create a scenario assessing infants' abilities to handle two representations belonging to the same object, Buttelman and colleagues (Buttelmann et al., 2015) created a scenario involving appearance-reality distinction that resembled the tasks developed for understanding beliefs about identity in preschoolers (discussed in the introduction of this thesis, Chapter 1.1.2; but see also Apperly & Robinson 1998;

Rakoczy, *et al*., 2015). In this paradigm, infants as well as an actor were shown an object by an experimenter. Then the experimenter demonstrated the object's 'real' identity (e.g. a rock that is in fact a sponge), which demonstration was sometimes seen by the actor (true belief trials) and sometimes it happened in her absence (false belief trials)[16]. Then this object was put high up on a shelf, and in the following test phase the actor acted as if she wanted to reach the object. While infants could not reach the original object either, two objects were revealed in front of them, one of which resembled the appearance, and the other the real identity of the original object. They measured which object infants hand to the actor as a response to this request for help. They found that infants reacted differently in the true and false belief scenarios; when the actor was unaware of the 'real' identity of the object (in false belief trials), they gave her the object that resembled the appearance of the original (e.g. the rock); whereas in true belief trials where the actor knew the real identity, they gave her the object corresponding to that (e.g., the sponge). This suggests that infants at 18-months can represent the aspect under which a person represents an object.

[16] In fact, while the authors refer to the two appearances as 'apparent' and 'real' identity of the objects; the objects differed with regards to what the 'real' identity would be. For instance, something that looks like a rock but is really a sponge, is in fact not a rock (as the rock-ness is defined by its material, which the sponge does not have); but a toy duck that is in fact a brush is *both* a toy duck and a brush, as the brush-ness doesn't take away the toy-duck-ness. Regardless of this distinction, however, infants would need to represent two aspects of the same object.

Infants in Buttelman *et al.*'s (2015) study understood when another person was knowledgeable or ignorant about the dual aspect of an object. Infants were aware of two attributes of this object; in true belief trials they ascribed both to the other person, and in false belief trials they correctly only ascribed representing one of these attributes to her. However, their scenario did not involve incorrect individuation of objects; as the infant and the actor represented the same number of objects.

The present study aimed to assess infants' ability to attribute to another person a false belief about object identity, when this involves the attribution of mistaken individuation. We tested 14-month-old infants, who were previously shown to be able to selectively vary the use of appearance information as the basis for individuation, based on their knowledge on whether two appearances may signal one object (Cacchione et al., 2013). Infants participated in scenarios where they were always aware of the dual appearance of an object, but we varied whether another person knows that the two appearances belong to the same object.

As in the paradigm of Experiments 3 & 4, infants and an actor first saw an object (in form A) being placed in an opaque box by an experimenter. Then in False Belief trials the actor left the room, and during her absence the experimenter transformed the object from form A to B, and put it back to the box. The actor then came back, searched in the box and retrieved the object in form B. In True Belief trials all events were identical, except the actor left *after* the transformation (and re-hiding), and in her absence nothing happened. Infants were then allowed to search in the box. Infants in both types of trials knew that there was only one object, which had two appearances; therefore at the time when they were allowed to search, the box was empty (i.e., all objects they knew were hidden, were also retrieved). In the False Belief

test trials the other person believed that one of the objects was still in the box (hereinafter, as in Chapter 3, we call these trials 'other believes =1', or OB-1); whereas in the True Belief trials the other person knew that there is no object in the box (hereinafter 'other believes =0', or OB0). Hence we reasoned that similarly to Experiments 3 & 4, longer search duration in the False Belief (OB-1) trials compared to the True Belief (OB0) trials would indicate that infants are sensitive to the social partner's beliefs in this scenario.

However, these inferences are rather complicated and infants may not always perform them even from their own perspective. In the study of Cacchione *et al.* (2013) not every infant behaved according to the predictions, i.e. to search longer in the no-pretraining group on switch trials, as there may be two objects; but search equally on switch and no-switch trials in the pretraining group who might have inferred that the two appearances belong to the same object. There overall approximately 70% of infants (9/12 in the no-pretraining and 8/12 in the pretraining group) showed the predicted behavior. While there may be many factors contributing to infants' behavior, it may be that not every infant made the necessary inferences. The possibility put forward in this thesis is that infants build on their cognitive systems responsible to represent the environment, to also represent others' mental states. Therefore we reasoned that infants' limitation to represent certain state of affairs may limit them in attributing such contents to others, hence overshadowing their ToM abilities.

Based on the above considerations we introduced a set of Baseline trials. In both Baseline trials infants saw initially form A being put in the box, and form B being

taken out right before they were allowed to search[17]. In fact, we used transformable objects, therefore form A and B belonged to the same object. Crucially, on unknown-switch baseline trials infants were not aware of this feature. However, on known-switch baseline trials they were shown that the object is transformed from A to B, and then put back in. We aimed to assess whether infants individuate on the basis of their knowledge about the two appearances. In unknown-switch trials infants should individuate two objects and hence should search longer compared to known-switch trials in which case they should know there is only one object. We predicted that specifically those infants, who made this distinction in baseline, should be able to represent other's beliefs involving mistaken individuation in test trials.

The present study therefore aimed to test whether 14-month-old infants who can selectively vary the use of appearance information as the basis for individuation based on their knowledge on whether two appearances signal one object; can also attribute to another person a false belief about object identity, when this belief is based on mistaken individuation.

4.3.1. Methods

Participants

Sixty-four 14-month-old infants participated (age range from 14;2 [Months; Days] to 15;3 mean age = 14;16); 27 were girls. Thirty-four additional infants were

[17] Two separate objects were used in the two baseline trials, but for simplicity we will refer to the two forms of the objects for either object as form A and B.

tested but not included in the analyses because they did not search in the Belief Trials (15), they fussed out (12), or due to parental interference or experimental error in the procedure (7).

Materials

We used a new white cardboard box (38*38*20 cm) with a 20*14 cm opening that was covered by an elastic cloth that prevented infants from seeing inside the box but enabled reaching into it. The box in this study looked identical to the one used in the previous studies, but did not have a secret compartment and was made slightly larger in order to have a large enough opening for the objects.

In Familiarization we used the same colorful whistles as before, with a ball inside that made a rattling sound. In the False / True Belief trials two transforming objects were used, one bird/hedgehog and one crab/frog. The two objects were identical in size (approximately 15x10x10 cm) and were made of textile, but the two objects and the two forms of each object differed in color (See Figure 4.4). For the Baseline trials two different transforming objects were used that were approximately half the size of the Belief trial objects, differed in color from each other and were different in color and material from the Belief trial objects. The Baseline trial objects had agentive features – arms, legs, and eyes – but did not resemble humans or any particular animal.

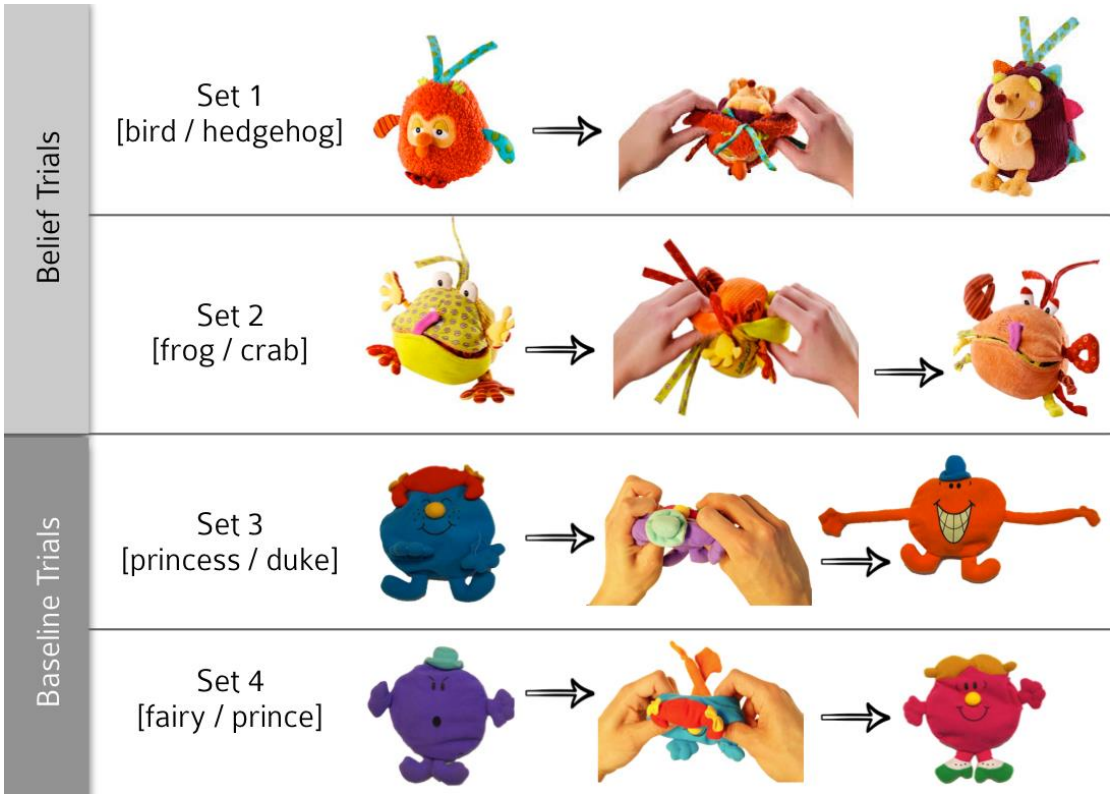

Figure 4.4. Transforming objects in Experiment 5.

Procedure

Infants sat in their parent's lap, at an 80 * 60 cm table with two experimenters sitting on the two longer sides of the table (as in Experiment 3 & 4). Due to relocation of the testing setup, for half the infants two cameras recorded the experiment from the infant's left and right angle; and for the second half a third camera recorded the infants from the front as well.

Each session began with two Familiarization trials, followed by four test trials: two Belief trials and two Baseline trials, with an additional Familiarization trial in between Belief trials and Baseline trials (see Fig. 4.5). The first two test trials

consisted of a False Belief (FB) and a True Belief (TB) trial. Each infant was tested on one FB and one TB trial; trial order was counterbalanced between participants. The second pair of test trials were Baseline trials in fixed order: an Unknown Transform and a Known Transform trial. The fixed order in Baseline served to avoid transfer effects from the Known Transform trial to the Unknown Transform trial (see Cacchione et al., 2013; where infants generalized a potential transformability feature from familiarization to test).

Familiarization trials

The first two Familiarization trials were identical to those in the previous studies: in the first E1 put in the toy whistle and retrieved it, and in the second she put in another and let the child retrieve it. The third familiarization came after the two Belief trials and served to re-engage children, as in the test trials their search was always unsuccessful (as the box was empty). In this third Familiarization E1 again put in a whistle and prompted the child to search for it.

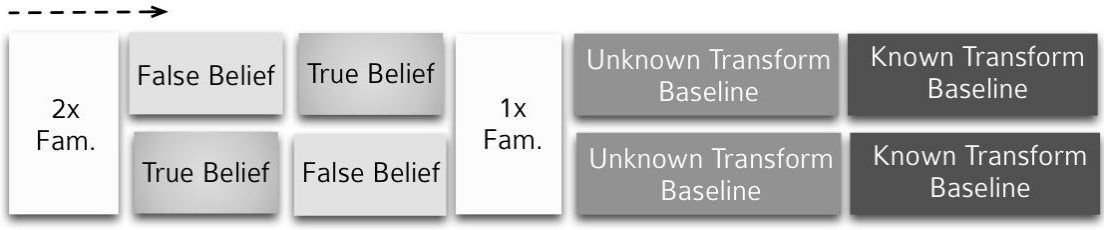

Figure 4.5. Trial structure of Experiment 5. Order of Test Trials (False Belief first or True Belief first) was counterbalanced between subjects; the order of baseline trials was fixed (Unknown Transform was always first).

False Belief Test trials

In Test trials always E2 took out the objects from her bag. This was done so to avoid any suspicion children might have that E1 has knowledge of the transforming

feature of the objects. E2 took out an object, put the it on the top of the box, pointed at it and said: "Look, a [bird/hedgehog/crab/frog]!". Then E1 repeated the label: "Oh, a [frog/ etc.]", and put it into the box. Following this, in FB trials E1 reached to her pocket and said: "Oh, my phone is ringing, I have to run out", and left the room. In her absence, E2 called the infant's attention saying "Look!", reached into the box, retrieved the toy and demonstrated the transformation. While she was turning the object inside out, she said "Look, do you see the [frog]? Let me show you something! The [frog] is also a [crab]! Do you see?" – and then put back the toy into the box.

Following this, E1 came back, sat down and performed the same searching action as in the first Familiarization trial; retrieved the toy, looked at it and said: "Oh, a crab! How nice!"; and then she put the toy away into a bag. Finally, she took a book from next to the box and said: "I have to look up something now", pushed the box in front of the infant and pretended to read for 15s. For this time E2 did not interact with the infant either. After the 15s E1 looked up, put down the book and said: "Ok, we are done". Then, to avoid "cumulating" beliefs, before continuing to the next Test trial E1 gave the toy to E2 to put it away, but before putting it away E2 showed it to E1 and said: "Look, let me show you something. Do you see the [crab]? The [crab] is also a [frog]". Hence, E2 always demonstrated the reverse transformation when E1 was already present. After this the second Test trial followed.

True Belief Test trials

TB trials were identical to FB trials with one difference: E1 left after E2 retrieved the toy and demonstrated the transformation, and came back right away. When E1 came back, she continued as in FB trials with retrieving the toy. When a TB trial was

first, E2 still did the demonstration before putting the toy away in the end of the trial, to match it to FB trials.

Unknown Transform Baseline Trials

In the first Baseline trials infants received a similar manipulation as one group in Cacchione *et al* (2013). In this trial, E1 took out an object from her bag, put it in front of the box, pointed at it and said: "Look, a [princess/prince/knight/fairy]". These namings did not match the figures as they did not resemble any known character, but served to name the two different forms of the objects. E1 then put in the object to the box, and after a brief pause she asked: "What's in the box? I am searching for it… searching… searching…". In the meantime, she transformed the object inside the box, and when she then retrieved it, it was in its changed form. E1 then said: "Oh, a [prince /etc]", naming the altered state, put away the toy and again pretended to read, after 'accidentally' pushing the box in front of the infant. After the 15s elapsed, E1 took out the same toy from her bag and showed the transformation to the infant, just as it was done in the Belief trials. This was done in order to demonstrate to infants that the box was in fact empty, and not to have carry-over effects in the next trial.

Known Transform Baseline Trials

This second Baseline trial was identical to the first Baseline, with the modification that E1 did not do the transformation inside the box, but retrieved the toy, showed the transformation, and put it back. Following this, she again pretended to read for 15 seconds while the box was in front of the infant.

Coding

As we tested whether infants attribute to another person a false belief based on individuation of objects relying on feature/identity information at the same age where this capacity has been found in their first-person inferences, we wanted to focus on infants who in our sample show evidence of such cognition. Therefore we aimed at a sample of 32 infants who in the baseline condition show a pattern of behavior consistent with the understanding that (i) if two objects seem to be of different kinds, then they are likely two different objects, and (ii) if there is evidence that one object can appear in two different forms, then this overwrites the inference from the first point and there is likely only one object. These two points are grasped by our baseline trials: in the unknown-transform trials infants should infer to be two objects (and hence one remaining at the time of search), whereas in the known-transform trials infants should come to the conclusion that the two forms belonged to the same object (and therefore none remains by the time of search). We determined a pre-set criterion we considered to grasp a large enough difference between search durations in the two baseline trials. Infants were considered to 'have passed' the baseline criterion if

$$Search_{unknown\_transform} \geq Search_{known\_transform} * 1.2$$

- that is, if search duration on the unknown transform trials was equal or higher than the search duration on the known transform trial multiplied by the constant of '1.2'; or if they did not search in the known-transform baseline trial, then a minimum of 500ms search duration in the unknown-transform baseline.

As we have no influence over which infant will pass this baseline criterion, we did not have a minimum number of infants who would *not* pass, rather we aimed at a

sample of 'passers' with equal number of infants from the two Orders of Test trials. Infants who are reported as non-passers are infants who (a) completed the study and searched to some degree in one or both baseline trials, however did not search longer in the unknown-transform baseline to fulfill our above described criterion, (b) completed the study but did not search at all at either of the baseline trials, or (c) did not complete all four trials (one infant did not search on the fourth – known transform baseline - trial).

4.3.2. Results

Baseline Trials.

In the final sample 32 infants were included who passed the baseline criterion, and 32 additional infants were tested who completed the two baseline trials but did not pass the baseline criterion. Infants who did not complete the two Belief trials (or

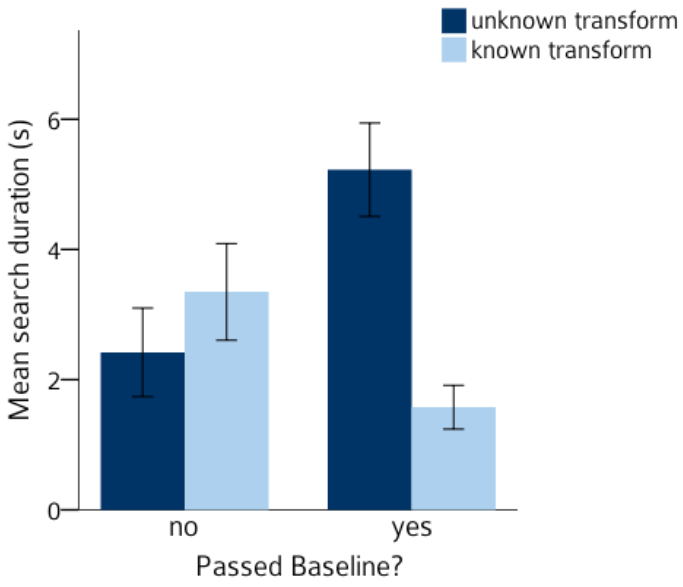

Figure 4.6. Mean search duration in Known Transform and Unknown Transform Baseline trials. Error bars depict +/- 1 Standard Error.

did not search on either of them) were excluded from all analyses. Search durations in Baseline trials are shown in Figure 4.6.

Belief Trials.

We first analyzed search duration during Belief Trials across all infants who completed the two Belief Trials. We conducted a 3-way mixed ANOVA with Order of Test (OB-1 first vs. OB0 first), Passing Baseline (Yes vs. No) as between-subjects variables, and Belief (OB-1 vs. OB0) as within-subjects factor. There was no overall effect of Belief ($F(1, 60)=0.196$, $p=.66$). Crucially, there was a significant Belief * Passing Baseline interaction ($F(1, 60)=6.891$, $p=.011$, partial $\eta^2=.103$; see Figure 4.7).

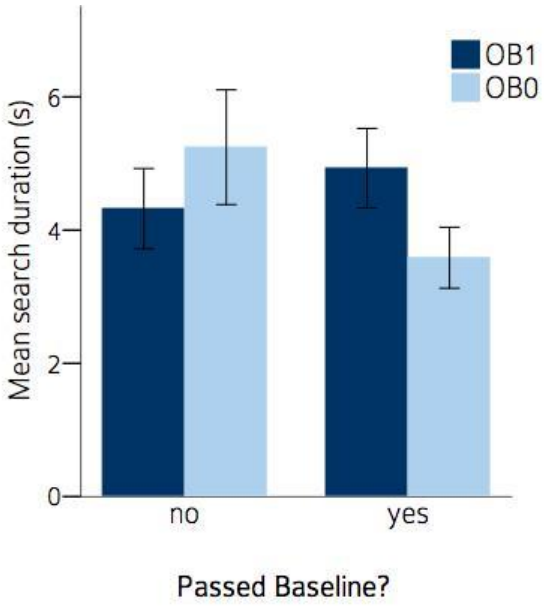

Figure 4.7. Mean search duration in 'other believes=1' (OB-1) and 'other believes=0' (OB0) trials in Experiment 5. Error bars depict +/- 1 Standard Error.

We then split infants into two groups based on whether they have passed the Baseline criterion, and in both groups ran a 2-way mixed ANOVA with Order of Test (OB-1 first vs. OB0 first) as between-subjects variable and Belief (OB-1 vs. OB0) as within-subjects factor. Infants who did not pass (n=32) did not show a significant

difference in search duration between OB-1 and OB0 trials (F(1, 30)=1.936, p=.174), with search durations somewhat longer in OB0 trials (M_{OB0}= 5.075, SE= .838) than in OB-1 trials (M_{OB-1}=4.118, SE= .594).

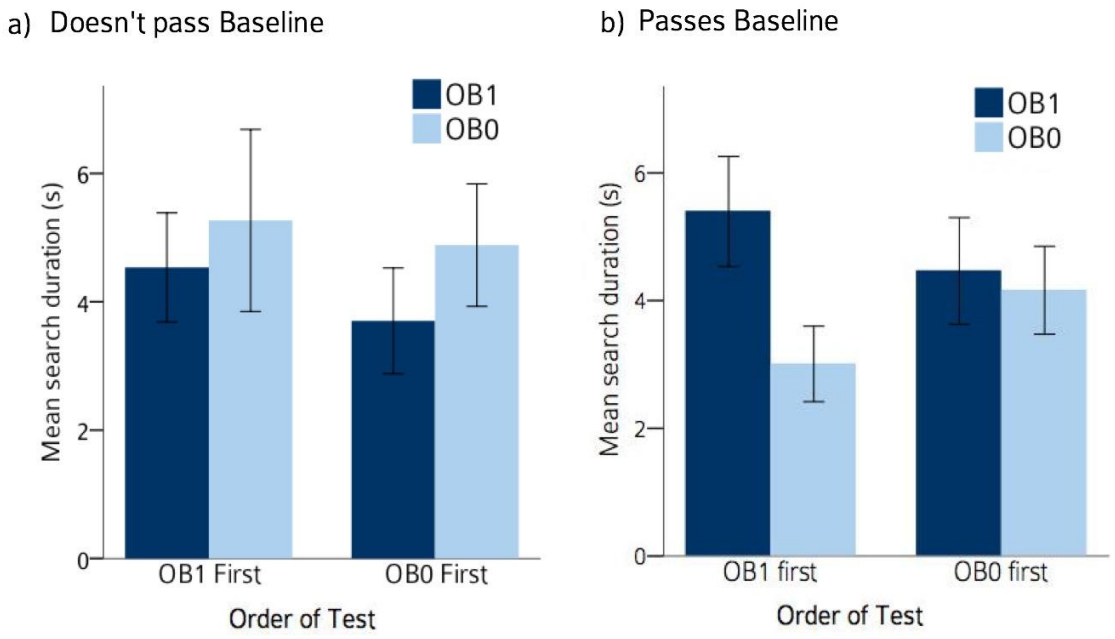

Figure 4.8. Mean search duration in Experiment 5, in 'other believes=1' (OB-1) and 'other believes=0' (OB0) trials in a) who don't pass Baseline and b) who pass the Baseline criterion. Error bars depict +/- 1 Standard Error.

Infants who passed the Baseline criterion (n=32) showed a significant main effect of Belief (F(1, 30)=6.106, p=.019, partial η^2=.169), with longer search durations in OB-1 trials (M_{OB-1}=4.93, SE= .6) compared to OB0 trials (M_{OB0}=3.586, SE= .454). Out of 32 infants 22 showed this pattern, and a binomial test indicated that this observed proportion of .69 was significantly higher than chance (p= .05, two-sided). There was also a marginally significant interaction between Belief and Order of Test (F(1,30)=, p= .065, partial η^2=.109; see Figure 4.8). When analyzed separately, there was a significant effect of Belief in the OB-1 first group (F(1,15)=11.921, p= .004, partial η^2=.443), but not in the OB0 first group (F(1,15)=0.13, p= .723, partial η^2=.009.

Additional analyses.

In order to assess the validity of our Baseline criterion, we checked the differences in search duration between the two Belief trials. The mean proportion in the "passed-baseline" group was 1.29, which is reasonably close to our criterion of 1.2 for the Baseline trials; and there were no infants who would not count as passing with a ratio of ~1.3 but count as passing with our ratio of 1:2.

We ran the same analysis with adding every infant to the 'passing' group who showed even minimal difference in search duration in the right direction (longer search in the unknown-transform Baseline), resulting in a sample size of n=39. The significant main effect of Belief remained (F(1, 37)=5.751, p=.022, partial η^2=.135), with longer search durations in OB-1 trials ($M_{OB\text{-}1}$=4.67, SE= .513) compared to OB0 trials (M_{OB0}=3.501, SE= .417). There was also a significant interaction between Belief and Order of Test (F(1,37)=, p= .018, partial η^2=.141). Overall, adding these 7 infants did not change our results, so we take as our main results the ones based on the pre-set criteria of the Baseline criterion.

Additionally, we tested the effect of the others' belief in the group of infants who searched on at least one baseline trial. In the initial analysis we categorized infants who did not search on either baseline trials as "non-passers", if they searched on at least one of the belief trials. While this is justified by the fact that these infants have already demonstrated their potential willingness and ability to search on the preceding belief trials (and hence the absence of search during baseline can be considered as a meaningful behavior); one could argue that infants might have been influenced by external factors such as shyness or distraction, and therefore the not

searching on either baseline trials leaves open the interpretation that they would make the necessary inference but not search on these trials for another reason.[18] Hence, we ran the 3-way mixed ANOVA with Order of Test (OB-1 first vs. OB0 first), Passing Baseline (Yes vs. No) as between-subjects variables, and Belief (OB-1 vs. OB0) as within-subjects factor; with only those infants as non-passers, who searched on at least one of the baseline trials. Results were highly similar to those of our main analysis: there was no overall effect of Belief ($F(1, 49)=0.008$, $p=.93$), and there was a significant Belief * Passing Baseline interaction ($F(1, 49)=7.289$, $p=.009$, partial $\eta^2=.129$). Similar analysis was not possible on the sample of 11 infants who did not search on the baseline trials, as there was no "Passing Baseline" factor; but a t-test showed no difference between OB-1 and OB0 trials ($t(10)=-0.451$, $p=.662$). Overall, treating infants who did not search on baseline trials as a separate group did not change our main results; if at all, effects seemed to be stronger if these infants were excluded.

[18] In fact, we do not wish to make a strong claim about non-passers. Infants could *not* pass the baseline for many reasons, and it is possible that some of the infants, who don't show the pattern of behavior in baseline trials, might still attribute to the other person the corresponding belief. Our claim is rather the reverse: we expect that those infants, who *do* pass the baseline, would tend to attribute the same inference to someone else too.

4.3.3. Discussion

In the present study we tested 14-month-old infants' ability to attribute to another person a false belief about object identity. We presented infants with scenarios involving dual-identity objects that could transform between two appearances. Each scenario started with an object in form A being put in an opaque box, and ended with the object in form B being taken out. Between trials we varied whether the infants knew about the transformability of the object (baseline trials) and whether another person was aware of it (test trials). We predicted that those infants who in baseline trials selectively behave based on the information they have about the object (whether they have seen it transform from form A to B, or not); would also successfully represent such scenarios from someone else's perspective. As in previous studies presented in this thesis (Experiment 4 in Chapter 3 and Experiment 5 in the present chapter), we measured such attributions indirectly, through assessing whether infants' search duration in the box varied depending on the other person's belief about the content of the box.

In test trials when infants were allowed to search, they had evidence that all objects (i.e., the one object that has been hidden) have been retrieved, and therefore there was nothing left in the box; but we varied between test trials what the other person knew about the content of the box. In the False Belief test trials the other person believed that one of the objects was still in the box, as she had not seen the object transform (OB-1 trials); whereas in the True Belief trials the other person knew that there is no object in the box (OB0 trials). Therefore we reasoned that similarly to Experiment 3 & 4, longer search duration in the False Belief (OB-1) trials

compared to the True Belief (OB0) trials would indicate that infants represent the other person's belief in this scenario. Crucially, as in Experiment 3 & 4, in none of the (baseline or test) trials was there actually something left to search for, therefore any difference in search duration between trials would reflect infants' inferences regarding the content of the box.

The baseline trials served to assess infants' own inferences. In the unknown-switch trial infants had a reason to believe that there are altogether two objects present, as they haven't seen the transformation; whereas in the known-switch they saw the transformation and hence they should have individuated only one object. As a consequence, in unknown-switch trials infants should have inferred that there is an object left in the box when they were allowed to search, contrary to the known-switch trials where they should think nothing remained. If as outlined above, infants could use the information about the transformation of the object for individuation, then this should manifest itself in longer search duration in unknown-switch trials. Based on the pattern of infants' search durations on baseline trials we therefore categorized infants into "passers" if they searched longer in the unknown-switch trial by a set constant.

This baseline criterion was set because we hypothesized that infants' performance with regard to mental state representations might be occasionally overshadowed by differences in representing the given content from their own perspective. Indeed, when looking at performance on the test trials, if all infants were analyzed together, there was no sign of infants representing the other person's belief. However, when we selected those infants who passed the baseline criterion, infants' behavior showed a main effect of the other person's belief. Specifically, infants

searched longer in OB-1 trials when the other person falsely believed that there is an object still in the box, than in OB0 trials when the other person witnessed the transformation and therefore she also knew the box was empty.

These results support the hypothesis put forward in this thesis that infants recruit their cognitive machinery that serves to represent the environment in the service of representing others' mental state contents. First, only those infants who showed a sign of making a particular distinction - whether feature/kind information about the object justifies individuating two separate objects – attributed this inference to the other person as well. Second, this attribution manifested itself again in a modulation of infants' own behavior by the other person's belief; which is in line with previous results (Kovács, Téglás, & Endress, 2010; and Chapters 2 & 3 of the present thesis) suggesting that infants' own representations and the ones they attribute to others have a common representational format.

The present findings suggest that 14-month old infants are able to understand aspectuality of belief representations, which is in contrast with predictions of the minimal-ToM view (Butterfill & Apperly, 2013). Infants' behavior in this experiment cannot be explained by positing registrations, i.e. relations between an agent and an object. If infants track such relations, that would predict attributing to the other person in both test trials a correct registration of the object, *outside* of the box (or discarding the registration overall, if they stop tracking it once it is taken out). In contrast, infants' behavior suggests that they successfully represented that the other person knew the objects *as* form A, but not B (e.g. she knew the object *as* a bird but not as a hedgehog), therefore when form B was taken out from the box, she must mistakenly individuate it as another object. Therefore they likely represented the one

object under two aspects from their own point of view, but ascribed to the other person one of these aspects as belonging to another object representation. Is it possible that infants did not represent the dual aspect of the object at all, and simply believed there are two objects? Two arguments rule out this option: first, in true belief trials they observed exactly the same transformation of the object and successfully attributed to the other person based on her perceptual access that she will represent one object; whereas being unable to represent the dual nature would lead them to behave in true belief trials like in false belief ones, i.e. to infer there is another object in the box. Second, it is those infants who in baseline trials demonstrated having made these inferences themselves, who also ascribed such representations to the other person.

How did infants in the present study conceptualize these objects? We refer to them as dual-identity objects, since if someone does not know that the two appearances belong to the same object, they might perceive it as two distinct identities of two individuated objects. It is worth noting that objects (or any entity) can have multiple aspects in many different ways. Some that were used in studies with kindergarteners were dual-function objects; such as a die, that can also function as an eraser (Apperly & Robinson, 1998). Similar to this were cases with a person who was presented by his name first, was then also referred to as his profession, e.g. Mr. Müller, who is also the firefighter (Perner et al., 2011), or the pen that is also a rattle (Rakoczy et al., 2015). These examples include object or persons who have two characteristics that they both possess at the same time (e.g. when Mr. Müller goes home he does not stop being a firefighter; or when he is at work he is still called Mr. Müller). Other studies have used objects that have one external feature that masks their true function: such as the rock that is in fact a sponge (Buttelmann et al., 2015).

These objects differ from the previously mentioned in that the function should overwrite their appearance. Finally, the inside-out turning objects such as in the present experiment, and previously used by others (Cacchione et al., 2013; Rakoczy et al., 2015), have two appearances that can turn into one another; therefore they have to alternate to be observable but nevertheless have both features at the same time (e.g. the hedgehog, if turned inside out becomes a bird; nevertheless it is *also* a hedgehog).

While these examples differ in how the different aspects are implemented; they share a common underlying attribute. Namely, they are able to display multiple features that can serve as basis for individuating an object; such as appearance, label, or function. Therefore each of them, in the absence of spatiotemporal information, will lead an observer who sees the two aspects sequentially to mistakenly individuate two objects instead of one. This is the consideration underlying aspectuality: the statement "Superman is in the kitchen" or "Clark Kent is in the kitchen" refer to the same external referent and therefore these terms can be used interchangeably; nevertheless "Mary believes Superman is in the kitchen" or "Mary believes Clark Kent is in the kitchen" cannot be used interchangeably. Therefore if both beliefs can be ascribed to Mary, but she is unaware that Clark Kent is Superman, then she will believe that *two* people are in the kitchen. This lack of substitutability is captured by the notion of referential opacity (Quine, 1961). Similarly, in the present study attributing to the other person "she believes the bird is in the box" cannot be substituted to "she believes the hedgehog is in the box". Therefore infants in the false belief trials likely attributed to the other person that the believes the bird to still be in there, if she has seen the bird being put in; as the fact that the hedgehog was taken out does not influence her belief about the bird.

Nevertheless, Rakoczy *et al.* (2015) point out alternative ways to conceive these objects, that according to their concerns may enable simpler solutions to such scenarios. They point out that the object could be perceived as (i) shifting identities over time, (ii) having neither A or B aspects but a third one that has both features, or (iii) form A containing also form B inside. These descriptions, while theoretically plausible, do not argue against children's or infants' understanding of aspectuality. First, *every* example of dual-aspect entities can be described in the above terms. Clark Kent sometimes changes into Superman, then back to Clark Kent - cf. point (i); therefore he is both and *neither* – cf. point (ii); and when he is Superman he is also Clark Kent "inside" – cf. point (iii). Second, infants in the present study would not have succeeded with strategies that do not involve metarepresenting the other person's representation of two objects. For instance, if infants represent the object as form A containing B; then when it is taken out, they should represent it as form B containing form A. The other person is unaware of this attribute of the object that just emerged in form B, but even if infants track this, by itself it does not lead infants to any inference about her belief. This information is only relevant insofar as infants use this information to attribute to the other person that she believes form A is a different object, therefore she mistakenly individuates two aspects of the same object into two different objects. This essentially fulfills understanding referential opacity: to understand that the statements "she believes the bird is in the box" cannot be substituted to "she believes the hedgehog is in the box"; therefore if she believes *both* statements, she believes *two* objects to be in the box.

In sum, the present study showed that 14-month-old infants can represent another person's false belief based on their mistaken individuation of one object into

two. This suggests that young infants' ToM abilities do not show some limitations proposed by recent theories, such as representing mental states about identity (Butterfill & Apperly, 2013). Moreover, this limitation should come from the operation of a system that is responsible for spontaneous belief tracking through the lifespan (Apperly & Butterfill, 2009); therefore in no age group should such computations take place spontaneously. Since we showed the attribution of beliefs about identity in a spontaneous-measure paradigm, this provides supportive evidence against this claim. Finally, the modulation of infants' own behavior by the other's belief, and the fact that only those infants successfully represented the belief of the other person who demonstrated such inferences from their own perspective; support the proposal that infants build on their cognitive apparatus enabling representing the environment, when they represent other's beliefs.

4.4. Conclusions

The present chapter investigated 14-month-old infants' abilities to ascribe false beliefs based on object individuation to others. In order to successfully represent someone else's mental states related to an object, one has to suppose that they individuate and successfully track that object. In the majority of cases this is possible based on spatiotemporal properties, such as observing the objects in two different spatial locations at the same time. Such tracking also enabled success in the experiment presented in Chapter 3 of this thesis: the objects were presented in a way that enabled individuation based on spatial location both from the other person's perspective, and from the infants' point of view. However, spatiotemporal information is not always available, therefore being able to rely on other types of

information for individuation in social contexts may be of essence. In the present chapter we explored cases where tracking spatiotemporal information is not enough to successfully predict individuation from someone else's point of view. We aimed to investigate whether infants can make correct inferences about someone else's belief, when this is based on the assumption that they individuate the objects based on feature/kind information. Experiment 4 showed that infants could successfully infer that someone can have a false belief based on a correct individuation of two objects that are of different appearance. Building on this, Experiment 5 showed that infants can ascribe to someone a false belief based on mistaken individuation of two appearances of one object into two objects; suggesting that infants at this age successfully represent others' beliefs about object identity.

Together, Experiments 3 -5 expand existing findings of infants' representations of other's beliefs. Most scenarios investigate representing one person's belief about one object, or occasionally multiple people's belief about one object. The present studies build on infants' rich abilities to represent and track objects around them. Here we explored whether infants can use their object-tracking abilities within the established limits of parallel individuation to represent other's beliefs. The findings are in line with the possibility that infants' ToM abilities are flexible and rich from early on, as they can entertain various types of belief contents, including beliefs about identity. Future research may investigate how such limits of the object-tracking system interact with infants' belief representations. Furthermore, as metarepresenting others' representations was suggested to be effortful and non-automatic; the flexibility of such processes should be investigated. Chapter 2 provided suggestive evidence that infants manipulate the represented belief contents on-line, as events are unfolding. It is however an open question, how flexible ToM processes are with

regard to radical revision or discarding already computed representations. In the following, Chapter 5 investigates this issue with a focus on goal-directed agency attribution.

Chapter 5

Flexibility of infants' agency- and goal attribution

5.1 Introduction

Representing others' mental states enables interpreting and predicting their actions, and planning one's own reactions accordingly. For this, the mindreading system has to be fast and flexible. Flexibility in the context of mindreading is often described with regard to the need to represent different types of mental states, or the need to reason about mental states with deliberation (Apperly & Butterfill, 2009; Butterfill & Apperly, 2013). However, another type of flexibility needed is the ability to incorporate new information into the mindreading process (Kovács, 2015). Broadly speaking, this can mean any type of event or information that one receives after the first event that triggered the computation of someone's mental state.

Many theories on ToM have addressed the circumstances under which mental state attribution may happen; for example, Leslie's theory predicts mental state computation to take place when there is an agent (Leslie et al., 2005); and other accounts also discuss automatic belief encoding triggered by the presence of an agent (Butterfill & Apperly, 2013; Kovács et al., 2010; Samson et al., 2010). However, they do not address whether once the process was set in motion, it could be stopped – for instance, whether mental state representations can be revised or discarded. If such computations happen in an automatic manner, they might be resistant to counterevidence once the representations have been formed.

While much research on ToM is concerned with belief understanding, many of these studies in fact rely on the assumption that infants understand goals, preferences, or desires. When infants' expectation of an actor's actions is assessed, this expectation is often based on the agent's goal to get (or a preference for-) a

specific target (e.g. Onishi & Baillargeon, 2005). Therefore assessing whether infants can update or discard belief representations has to follow testing first whether they can perform these operations on goal or preference attributions. Otherwise if infants were found not to expect an agent to act towards a goal based on her belief, this could be due to failure of updating the belief representation, the goal representation, or both.

To take the first step towards answering these questions, this chapter therefore aims to investigate the flexibility of infants' mental state calculations with regard to goal attribution. Infants from a young age understand goal-directed actions (Csibra, 2008; Gergely & Csibra, 2003; Luo & Baillargeon, 2005; Woodward, 1998). In her seminal study, Woodward (1998) habituated infants to an event in which a hand repeatedly grasped one of two toys. After habituation, the location of the toys was swapped and in test the hand grasped either the old object in the new location or the new object in the old location. Results showed that the infants were surprised at the events in which the hand grasped the new toy, suggesting that infants at 5–6 months of age encoded the goal of the action (the old object) and expected the hand to act accordingly.

Infants have been since found to attribute goals to a variety of non-human agents as well. However, such attributions depend on the presence of the right cues to trigger such interpretation (Daum, Attig, Gunawan, Prinz, & Gredebäck, 2012; Gergely, Nádasdy, Csibra, & Biro, 1995; Hernik & Southgate, 2012; Kuhlmeier et al., 2015; Shimizu & Johnson, 2004). But do infants take into account information about such cues if they encounter them at a later time point? The present set of studies asked whether goal-directed agency attribution can be revoked, if infants face

information that question the validity of the attribution. We aimed to present a scene in a way that would initially trigger goal-directed interpretation of an entity's actions, only then to reveal another part of the scene suggesting that the basis of the original interpretation was mistaken. We reasoned that manipulating infants' impression whether an entity moves in a self-propelled way could influence whether they take its actions to be goal-directed.

Various studies have investigated the characteristics an agent has to display (or a scene has to contain) in order for its actions to be interpreted as goal-directed. For instance, Biro & Leslie (2007) found that self-propelledness (i.e. self-initiated movement), together with a salient action-effect elicited attribution of goal-directedness in 9- and 12 month old infants[19]. On the other hand, others have found that self-propelledness does not have to be demonstrated, if the movement otherwise can be perceived as rational action; that is, achieving a certain goal the most efficient way, given the situational constraints (Csibra, Gergely, Biro, Koos, & Brockbank, 1999). It may be that when an entity displays rationality of action, then giving infants evidence that it is self-propelled is not necessary in order to interpret its actions in a goal-directed manner, but in the absence of such cues self-propelledness is needed for goal attribution. In line with this, in their study Luo & Baillargeon (2005) found that if an agent moved in a possibly self-propelled manner (without displaying rationality of action) and had a small handle attached to it, infants saw its approach of a target as goal-directed; however in another group of infants where the handle seemed to end outside the infants' view (e.g. outside the stage), and therefore the

[19] However, for 6-month-olds equifinality of action was also necessary to attribute goal-directedness.

agent was *possibly* moved by an external force; infants did not attribute goal-directed agency.

Additionally, even if they are not given evidence that an agent moves by itself (e.g. in Csibra et al., 1999), infants might *assume* that the agent is self-propelled. The combination of two findings supports this possibility. First, Saxe and colleagues (Saxe, Tzelnic, & Carey, 2007) found that if infants were first familiarized with an inert object (a beanbag), and then they saw the object 'fly' across a stage; they inferred that in the direction where it possibly flew from, there must be a causal agent (e.g. a hand) outside of the stage, and not an inert block. Relatedly, in the experiment of Csibra *et al.* (1999) the agent entered the scene when it was already in movement, and then changed its movement according to the situational constraints, to reach a target. As Saxe *et al.*'s findings suggest that infants assume that movements require a causal agent; the infants in Csibra *et al.* either could have assumed that the agent itself is self-propelled, or they might have inferred that there must be *another* agent hidden who caused its movement. However, the latter possibility is problematic, because if infants see the movement of an entity to be caused by an external factor, they do not seem to interpret its actions as goal-directed (as shown by the long-handle condition in Luo & Baillargeon, 2005).

The present study built on these findings, by manipulating information that may guide infants' perception of self-propelledness of an entity. Initially, we presented infants with an event that showed an inanimate object that moved in a seemingly self-propelled way. Results from the short-handle group in Luo & Baillargeon (2005) suggest that self-propelledness may elicit the attribution of goal-directedness even in the absence of other indicators (such as rationality of action). However, later we

presented infants with cues that meant to either confirm, or disconfirm the self-propelledness of the agent. If an entity is initially shown *not* to be self-propelled, and the scene does not contain any other indication of goal-directedness, then infants should not understand the actions of the entity as goal-directed. Through testing how infants react to such information after they have already ascribed goal-directedness to an entity, we aimed to assess the flexibility of integrating information into infants' existing representations.

As discussed above, there is ample evidence that infants (under specific circumstances) expect an agent to continue to act based on its goal. However, in most of these cases infants' expectation is measured right after observing the agent's object-directed actions. It is therefore an open question whether infants can successfully incorporate additional information in an already formed representation. Such information can be confirmatory, and simply needed to be added to their representation of the agent. However, occasionally, new information might require revising or discarding an existing representation. Based on the above considerations we reasoned that evidence suggesting that an agent is in fact not self-propelled, and thus likely not an agent with goals, might require such radical revision.

In sum, the current set of studies aimed to investigate whether infants, after observing events that would induce goal-directed agency attributions, can *keep* this attribution after integrating new information into their representation; or *discard* such an attribution, if they receive additional evidence suggesting that their original inference was mistaken.

5.2. Experiment 6

To address these questions the present studies adapted the paradigm of Luo & Baillargeon (Luo & Baillargeon, 2005). There, in Experiment 1 they first established that 5-month-old infants in live demonstration expect a box, whose only agentive feature was self-propelledness, to act in a goal-directed manner (i.e. infants looked longer when the box in test phase approached the other target as before). Then in Experiment 2 they introduced a small manipulation, to test the contribution of self-propelledness to infants' interpretation of the events. They added a short (Agent condition) or long (Non-agent condition) handle to the same box as before. As described above, infants who saw the short-handle box expected it to act in a goal-directed manner; whereas the long-handle group did not have an expectation on how the box should act.

In the present study we used 3D animations. The sequence of events started as in Experiment 1 of Luo & Baillargeon (2005). Specifically, in an orientation phase we first showed a rectangle-shaped object (henceforth, the 'agent') alone that moved back and forth on a flat surface (observed from side view) on the screen. Then in familiarization two objects appeared on the screen, and the agent repeatedly approached one of them. This was followed by our crucial manipulation; we introduced a second orientation phase, where the bottom of the screen was revealed: what seemed before as a uniform surface supporting the agent, were in fact two occluders that could open in a curtain-like manner. This allowed us to potentially add various types of information to the existing scene. When the occluders opened, this revealed that in fact there was a handle attached to the agent. The handle, following

Experiment 2 of Luo & Baillargeon (2005), could be either long or short. In our Agent condition the handle was short, as previously it was found to be compatible with attribution of goal-directedness. In the Non-Agent condition the handle was long, and reaching to the bottom of the scene, as if it ended outside of the screen (however, unlike in the setup of Luo & Baillargeon, here the handle reaches downwards, whereas in their setup it reached sideways to the left or right of the scene). This was then followed by a test phase were we measured whether infants expect the agent to continue to act goal-directedly.

Since previously infants were found to react differently to the actions of a box with a long and short handle, this allowed us to compare it to our scenario where they were only later presented with this information. Specifically, if they can integrate this new information into their existing representation, then after seeing the short handle they should continue to perceive it as a self-propelled agent. A failure to do so would result in the lack of expectation on the agent's behavior in test phase. In contrast, if seeing the long handle causes infants to flexibly revise their interpretation of the scene, then they should not perceive the agent as self-propelled anymore and therefore not expect it to act goal-directedly in test. However, if infants' representations are inflexible and cannot be modified or discarded; then they would expect the agent to act in a goal-directed manner also in the long handle condition.

In the first study only the Agent condition was implemented, to validate the paradigm. If infants perceive the movement of the rectangle as goal-directed action, and then keep this attribution after an additional event that provides them with additional information that however does not defy the agent-status of the rectangle; then they should expect it to continue to act in a goal-directed manner. This would

manifest itself in the so-called 'Woodward-effect' (after the effect found in Woodward, 1998): after a swap of the objects' location, longer looking during events where the agent approaches the new goal in the old location (which we will refer to as Inconsistent choice events) than during events when the agent approaches its old target in the new location (Consistent choice events).

The goal of the first experiment was to test whether infants would perceive an animated self-propelled agent to act in a goal-directed manner in a modified Woodward-paradigm. Our crucial manipulation was that after familiarization (that served the purpose of goal-induction) we added an extra 'orientation' phase; which enabled providing extra information that should prompt infants to possibly update but nevertheless keep their representation of the agent.

5.2.1. Methods

Participants

Twenty-four full-term 9-month-old infants were tested (11 girls), age range from 8;15 [Months; Days] to 9;15, mean age = 8;27. Eight additional infants were tested but not included in the final analysis due to technical error (2), because the infant was fussy or cried (5) or the infant reached maximum looking criterion in all trials (1). Parents signed informed consent prior to participation, and the study was approved by the Hungarian Psychological Ethical Committee (EPKEB).

Procedure

Infants sat on their parent's lap and watched short 3D animations on a 90*50 screen in a dimly lit room. Parents were instructed to keep their eyes closed throughout the experiment and not to communicate with the infants. A camera recorded infants' looking behavior, which was also projected in the outside room to a monitor where the experimenter was standing.

Videos were presented with PsyScope X B77 (open source) software. During infant-controlled periods of the study, the experimenter assessed infants' looking through the monitor, and pressed a mouse button when the infant was looking, and released when the infant looked away. Trials ended automatically after the set time elapsed (see below). Infants saw 10 trials in the following order: 2 Orientation trials ('Orientation I.'), 4 Familiarization trials, 1 Orientation trial ('Orientation II.'), 1 Display trial, and two Test trials.

Stimuli

We presented infants a sequence of short movies (see Figure 5.1) showing 3D animations created in Blender animation program (open source, version 2.74, 2015-mar-31). Each movie was preceded by an attention getter animation with a blue flower on a grey background that either rotated or increased/decreased its size, accompanied by a brief sound effect. The movies contained sound effects accompanying the movements, in order to keep infants' attention.

Orientation I.

In the first Orientation phase infants saw two trials where a green rectangle moved back and forth on the screen. In one trial the rectangle started from the middle of the screen, and in 2s moved to the right side, then turned in 2s, waited 1s, moved to the left side in 2s, turned around in 2s, waited 1s, then went back to the center in 1 s and waited 1s. The start and the turning events were accompanied by little sound effects to keep infants' attention on the screen. The video was looped until infants looked away for 2 consecutive seconds or looked for 15 cumulative seconds. In the second trial the same events were shown, with the difference that the rectangle's first movement started in the other direction.

Familiarization

In Familiarization the same green rectangle was shown, with two objects (a blue cup and a pink cone) on the left and right side of the screen. In the four familiarization trials the rectangle first stood still for 1s, then started moving towards one of the objects, approached it in 2s and stood still next to the object for 5s. The start of the trial was accompanied by a short sound effect to grab infants' attention. In all four trials the rectangle always approached the same target. The trials lasted 8s altogether, or ended earlier if the infant looked away for 2 consecutive seconds after the approach of target.

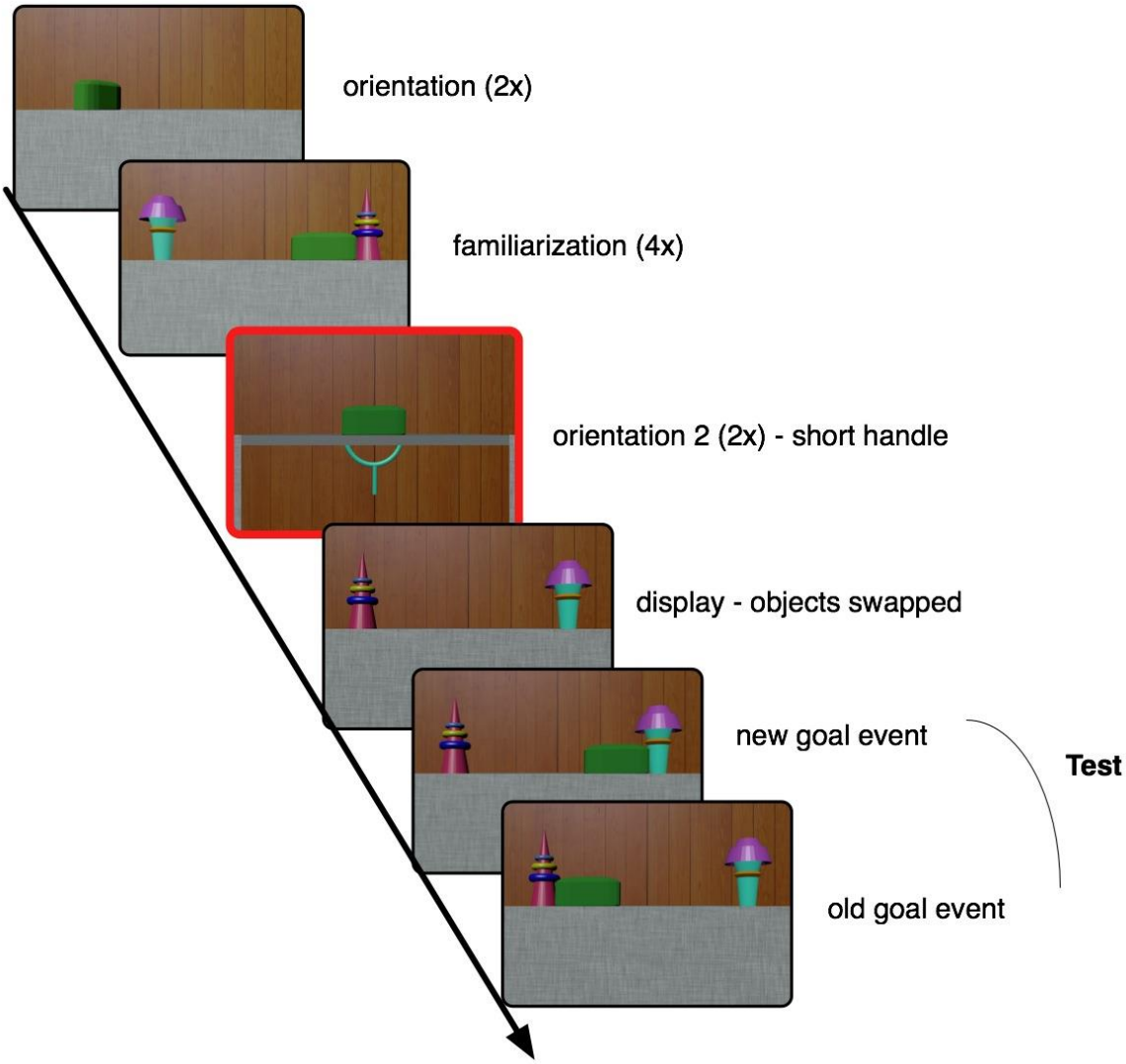

Figure 5.1. Event sequence in Experiment 6.

Orientation II.

Following Familiarization, infants saw another Orientation event. First, the green rectangle was visible again and the objects were absent. Then what seemed so far to be a uniform solid surface supporting the objects, but was in fact two occluders with just a thin surface below the objects, opened in the middle and started moving apart for 5s with its final position almost fully opened and only the edge of the occluders visible on both sides of the screen. When the occluders moved apart this revealed the

lower side of the screen and the bottom part of the rectangle object became visible. This bottom part was a plier-like part attached to the rectangle form below, and a short handle reaching down from the plier. Following Luo & Baillargeon (2005), the short handle should not interfere with agency attribution based on self-propelledness.

Then the rectangle and the bottom part started moving, first elevating and lowering back (2s). Then the same left-right-turning movement was shown as in Orientation 1 with the same timing of events, but this time, as the occluders were still open, the bottom part was also visible and moved together with the green rectangle. This video was again looped until infants looked away for 2 consecutive seconds or looked for 15 cumulative seconds.

Display

Before the two test trials were presented, a 5 s display of the two objects in the swapped position was shown, and the rectangle was absent.

Test

Finally, two Test events followed with the objects remaining in the swapped position, the rectangle present, and the occluders closed again. One trial showed an Inconsistent choice (IC) where the rectangle approached the new target object in the old location, and the other a Consistent choice (C) where the rectangle approached the old target in the new location. Timing of events was the same as in the

familiarization trials; but the scene remained present until the infant looked away for 2 consecutive seconds, or looked for 30 cumulative seconds.

The arrangement of the two target objects, the object of choice in familiarization, and the order of test events were fully counterbalanced across infants. The direction of first movement in the first orientation trial was randomized. Infants' looking time was measured on-line during test trials, and coded off-line with PsyCode coding software (open source, version 1.1). One-third of the videos were coded offline by one of two second coders to assess reliability of coding; inter-rater agreement was Pearson's r= .926, p< .001.

5.2.2. Results

Infants' looking time was analyzed in Consistent (C) and Inconsistent (IC) choice trials to assess whether infants expected the agent to continue to approach its goal despite the location change. Preliminary analyses showed no effect of the infant's sex, arrangement of objects (cup left vs. cup right), or target object (cup vs. cone), therefore these were omitted from subsequent analyses.

We analyzed looking data in a 2-way mixed ANOVA with Choice (IC vs. C) as within-subject variable and Order of Test (C first vs. IC first) as between-subjects factor. There was no main effect of Choice (F(1,22)= 0.008, p= .929) and no other main effect or interaction, with infants looking equally long during Consistent events (M_C=9.092, SD= 9.195) as in Inconsistent events (MI_C=9.282, SD= 9.66; see Figure 5.2).

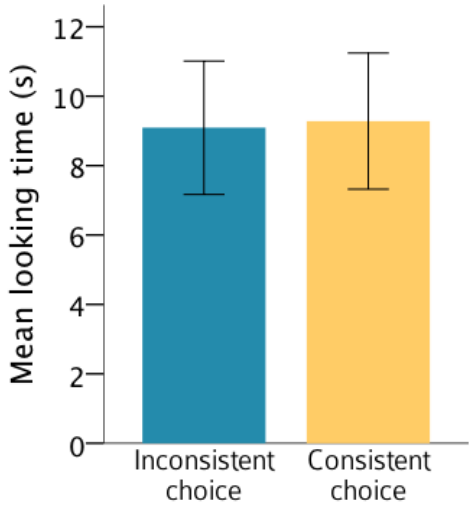

Figure 5.2. Mean looking times in Experiment 6. Error bars show +/- 1 Standard Error.

5.2.3. Discussion

We measured infants' expectation of an agent's behavior. A rectangle first demonstrated self-propelledness and approached repeatedly the same target object out of two. After goal induction an opening of occluders revealed to infants that in fact the agent had additional features that were not seen before. Following this, the location of the objects was swapped, and infants were shown two kinds of test events: in one the agent approached the same target as before (which we refer to Consistent choice), in the other the agent approached the new target (Inconsistent choice events). Infants' looking times were measured during the two kinds of test events, to assess whether infants are surprised if the agent chooses the new target and not act according to its previously demonstrated goal; which would be suggested by longer looking times to the Inconsistent choice events.

In order to expect the agent to continue to act on its goal, infants had to (i) encode the rectangle initially as a goal-directed agent, (ii) encode and remember the

agent's goal from familiarization phase, (iii) perceive the agent as continuous through the sequence of events, (iv) integrate the additional feature information into their representation of the agent (or alternatively, ignore it), and (v) not overwrite the agency- or goal- attribution due to the extra information.

Results show that infants in test phase did not expect the agent to approach its target object after the object locations were swapped, as they looked equally long to the two kinds of events. Infants' lack of surprise can be due to a variety of reasons, as it might be the result of failure at any points (i)-(v) above. First, it is unclear whether self-propelledness is a strong enough agency cue, as some studies suggest that self-initiated motion is neither necessary nor sufficient to elicit goal-attribution (Csibra et al., 1999; Schlottmann & Ray, 2010; Shimizu & Johnson, 2004); while others, however found it effective, though with 3D objects (Luo & Baillargeon, 2005). Second, even if they have perceived the action as goal-directed, infants might not have had enough time to encode the target object. Some others have used longer times during familiarization to allow infants explore and encode the scene, up to 30s total after the agent has reached the target (Hernik & Southgate, 2012). It is therefore possible that infants simply needed more time to later remember the target of the agent's actions.

It is also possible that due to the additional orientation event infants' memory of the familiarization phase faded enough not to produce strong expectations about the agent's behavior; or that they did not see the events as continuous but rather as independent events, which might have disrupted their interpretation of events. Alternatively, infants' failure to expect the agent to approach its goal might be due to *unsuccessful integration* of the new information into their representation of the agent (whereby the additional information interferes with infants' existing representation),

or possibly *successful revision* of agency attribution if for some reason they perceived the handle as an external source of energy (e.g. seeing even the short handle triggered associations of being moved by someone else[20]).

Finally, since infants received altogether 7 infant-controlled events before they reached test phase (2 orientation, then 4 familiarization, then one more orientation); infants might have simply lost interest in the test events. In addition, the many infant-controlled trials also introduced extra variability in infants' exposure to the relevant events. Based on these considerations, in Experiment 2 we implemented changes that targeted these possible issues.

5.3. Experiment 7

In Experiment 7 we aimed to address possible factors that may contribute to infants' lack of expectation that the agent would continue to approach its target object in test. Our primary goal was to implement a modification that enables us to assess whether infants initially perceived the actions in familiarization as goal-directed. If they did, the lack of expectation of goal-directedness may have been due to memory factors, to the lack of integration of new information with the previously received, or the result of the revision of agency attribution.

[20] While in the study of Luo & Baillargeon infants did not consider the short handle as signaling lack of self-propelledness or control over its movement, while we carefully matched our scenes to their descriptions, we cannot exclude that there are some differences in the appearance between our agent and theirs.

We therefore modified the original design by adding an extra test phase right after familiarization. As such, it is a closer replication of the Woodward-paradigm (Woodward, 1998), because the test phase directly follows familiarization. This allowed us to assess whether infants perceived the events in familiarization as goal-directed. Furthermore, we (i) added extra time for encoding the target object in familiarization trials, (ii) set infant-controlled trials only in familiarization but not in orientation to reduce the number of trials where infants need to disengage from the screen, (iii) added a left-right turning behavior in the beginning of familiarization trials to increase perceived selectivity in the agent's action, (iv) introduced familiar target objects which may be remembered more easily, (v) changed the size of the agent to make it more proportional to the overall scene as well as the target objects; and (vi) made the movement of the agent slightly more dynamic (somewhat faster and less uniform speed during the movement).

In order to aid infants in perceiving the sequence of events as continuous (i) the occluders did not open fully to show infants what it looks like when the handle is moving while partially occluded (therefore showing that its movement is likely this way when it is *completely* occluded, i.e. in familiarization, as well as in test phase), and (ii) in the end of the second orientation phase the occluders were shown as they were closing; to aid infants in their inference that the scenes they see throughout is belong to one event sequence.

We tested infants in two conditions: the Agent and Non-agent condition. These two differed only in their second orientation phase. First infants received two orientation trials that showed the rectangle move in a self-propelled way. This was

followed by 4 familiarization trials showing the agent repeatedly approach the same target object. After this, infants received the first test trial pair in order to assess whether infants perceived the actions as goal-directed; which included a Consistent choice (the agent approached the same object as in familiarization) and an Inconsistent choice trial (the agent approached the other object). When the first test phase ended, in the second orientation trial infants in both conditions saw the same events, with one crucial difference. In the Agent condition, as in Experiment 6, the opening occluders revealed a short handle below the rectangle. In the Non-agent condition the handle was long and reached 'outside' the screen (i.e. its end was not visible). Finally, infants received the second test trial pair, which was identical to the first one; in order to see how infants perceive the agent's actions after the second orientation phase, depending on the experimental condition (Agent vs. Non-agent).

We predicted that if infants indeed perceive the events in familiarization as goal-directed action; then in the first test phase all infants, regardless of experimental condition (as both groups observed the same events until this point) should expect the agent to approach its old target, and hence look longer during Inconsistent choice trials than in Consistent choice trials. Given the potential success of attribution of goal-directedness (as measured in the first test phase), in the second test phase infants' looking pattern might differ between conditions. In the agency condition we predicted a similar pattern as in the first test phase, if infants perceived the events as continuous and their memories or event tracking weren't disrupted otherwise. In the non-agent condition if infants successfully revise their interpretation of the events, we should see the effect of the first test phase go away. Infants should either not show any difference between trials or they should expect it to go to the old direction, as mechanical devices might do. However, if infants cannot revise their interpretation in

the Non-agent condition, then they would show the same looking pattern as in the first test phase.

In sum, the present study served two purposes. First, to assess whether infants can perceive the self-propelled motion of an animated object as goal-directed. Second, if they do, we aimed to test whether they can (i) keep this attribution if they face new information that should *not* change this attribution, and (ii) discard this attribution if they receive evidence that they *should* change their interpretation of previous events.

5.3.1. Methods

Participants

Forty-eight full-term 9-month-old infants (24 girls) were tested in two experimental conditions (24 per condition), age range from 8;15 [Months; Days] to 9;15, mean age = 8;27. Thirty additional infants were tested but not included in the final analysis due to parental interference (3), experimenter error (6), or because the infant was fussy or cried (20) or reached maximum looking criterion in all trials (1). Parents signed informed consent prior to participation, and the study was approved by the Hungarian Psychological Ethical Committee (EPKEB).

Procedure

The experimental setup was identical to Experiment 6, with the modification that infants received 2 test trial pairs. Infants saw 12 trials in the following order: 2

Orientation trials (Orientation I), 4 Familiarization trials, 2 Test trials (Test Pair 1) 1 Orientation trial (Orientation II), 1 Display trial, and two Test trials (Test Pair 2).

Stimuli

Stimuli were identical as in Experiment 6, with the modification that the target objects were exchanged to familiar ones (a cup and a ball, see Figure 5.3), in order to aid infants' memory in remembering the goal object. In addition, the size relation between the 'agent' and the objects changed: while in Experiment 6 the rectangle was small relative to the targets, in Experiment 2 the rectangle was approximately as tall as the targets, and wider than them. This modification was implemented because in several other studies with human agents (Cannon & Woodward, 2012; Woodward, 1998) or non-human agents (Csibra et al., 1999; Hernik & Southgate, 2012) where infants were found to successfully attribute goal-directedness, the agent is comparable in size, or even bigger than the goal objects. In addition, both the agent and the objects were made slightly larger to make them more salient and recognizable.

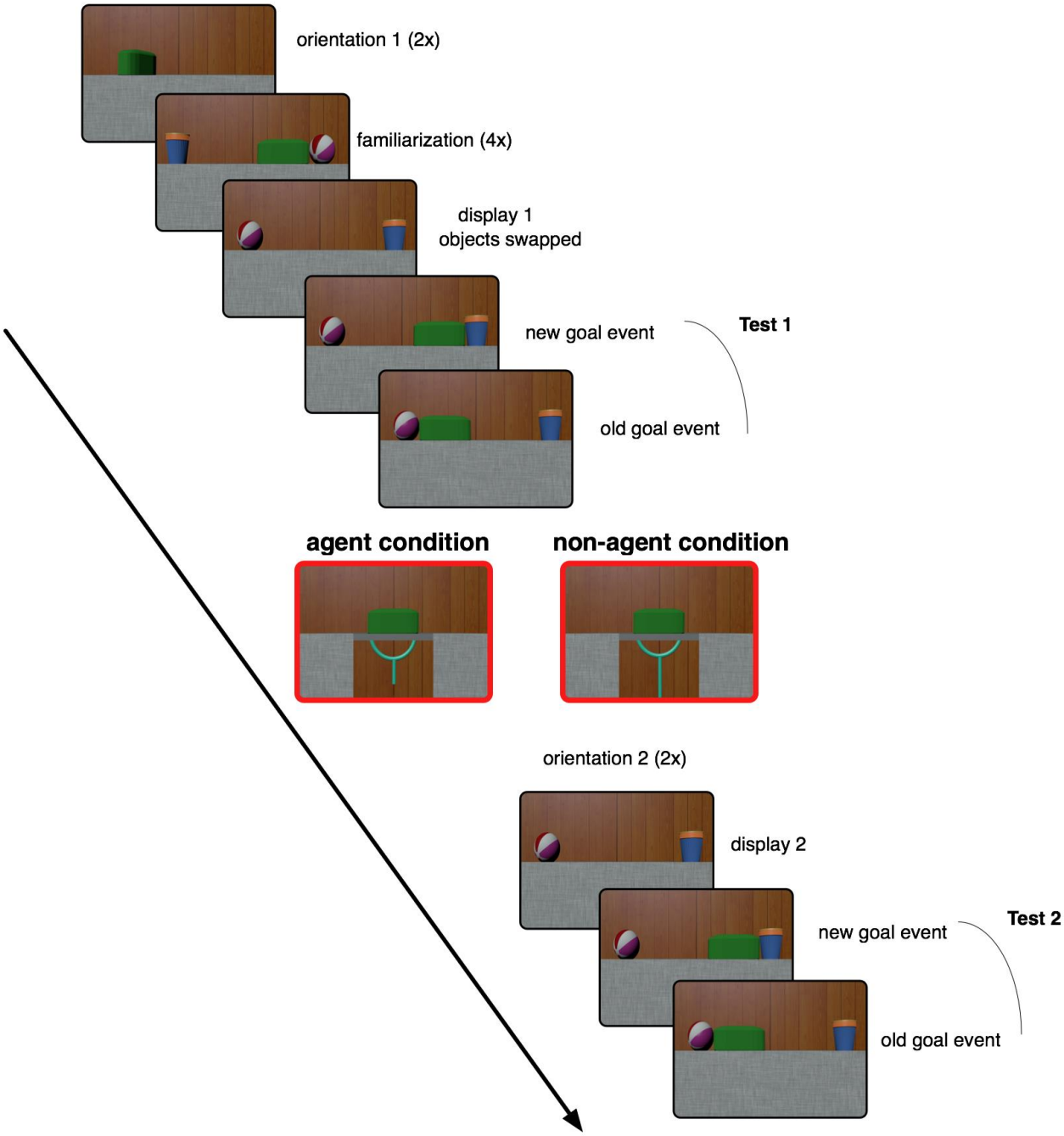

Figure 5.3. Event sequence of Experiment 7

Orientation I.

In the first Orientation phase infants saw two trials where the green rectangle moved back and forth on the screen. The movement of the rectangle was similar to Experiment 6, but was slightly faster to convey more natural movement. In one trial the rectangle started from the middle of the screen, and in 2s moved to the right side, then turned in 2s, moved to the left side in 2s, turned in 2s, then went back to the center in 2s; resulting in a total 10s video. The start and the turning events were accompanied by little sound effects to keep infants' attention on the screen. In the second trial the same events were shown, with the difference that the rectangle's first movement started in the other direction. The orientation videos, unlike in Experiment 6, were not looped and played to the end of the video, at which point the trial ended.

Familiarization

As in Experiment 6, in Familiarization the same green rectangle was shown, with two objects (a blue cup and a striped ball) on the left and right side of the screen. We introduced in the beginning of the trials an orienting-like behavior of the rectangle. This was done in order to convey that the agent can selectively vary its actions, which has been shown to contribute to the attribution of goal-directedness (Biro & Leslie, 2007; Csibra, 2008). While in most other cases they varied the route through which the agent reaches its goal; possibly other indicators of showing that the behavior can vary in direction might convey a similar notion.

In the four familiarization trials the rectangle first stood still for 1s with its shorter side towards the viewer's angle; and turned 45° right-left in 3-3s respectively; then turned 180° to one side (counterbalanced across infants) in 1s, then waited 1s, and then approached the object on that side in 2s and stood still next to the object until the infant looked away for 2 consecutive seconds or watched a cumulative 15 seconds. The start of the trial as well as the start of movement was accompanied by a short sound effect to grab infants' attention, and to familiarize infants with a particular sound signaling the onset of movement of the rectangle. In all four trials the rectangle always approached the same target.

Display 1

Before the first two test trials were presented, a 5s display of the two objects in the swapped position was shown, and the rectangle was absent.

Test 1

Two Test events followed with the objects remaining in the swapped position, and the rectangle present. One trial showed an Inconsistent choice (IC) where the rectangle approached the new target object in the old location, and the other a Consistent choice (C) where the rectangle approached the old target in the new location. Unlike familiarization trials, the rectangle started at the 180° position, not to cue infants about the direction of movement. After approaching the target, the scene remained present until the infant looked away for 2 consecutive seconds, or looked for 30 cumulative seconds.

Orientation II.

Following Familiarization, infants saw another Orientation event. This was identical to that of Experiment 6, with two modifications. First, the occluders did not slide out as far; therefore during the movement of the rectangle the handle became partially occluded. This was implemented to aid infants' understanding of how the scene before related to the present event (i.e. that before the handle moved behind full occlusion). Additionally, in the end of the scene (after the movement of the rectangle) the occluders closed, again to help infants perceive the events as continuous (as otherwise it might be less clear how the next scene with the occluders closed relates to the preceding orientation event).

Display 2 & Test 2

The Display 2 and Test 2 events were identical to the first respective events, with the object remaining on the same side as in the first events (and therefore swapped compared to familiarization).

The arrangement of the two target objects, the object of choice in familiarization, and the order of test events were fully counterbalanced across infants. The direction of first turn (left vs. right) in the first familiarization trial was randomized.

5.3.2. Results

Infants' looking time was analyzed in two test phases, each including one Consistent (C) and one Inconsistent (IC) choice trial. Preliminary analyses showed no effect of the infant's sex, the arrangement of objects (cup left vs. cup right), or target object (cup vs. cone), therefore these were omitted from subsequent analyses.

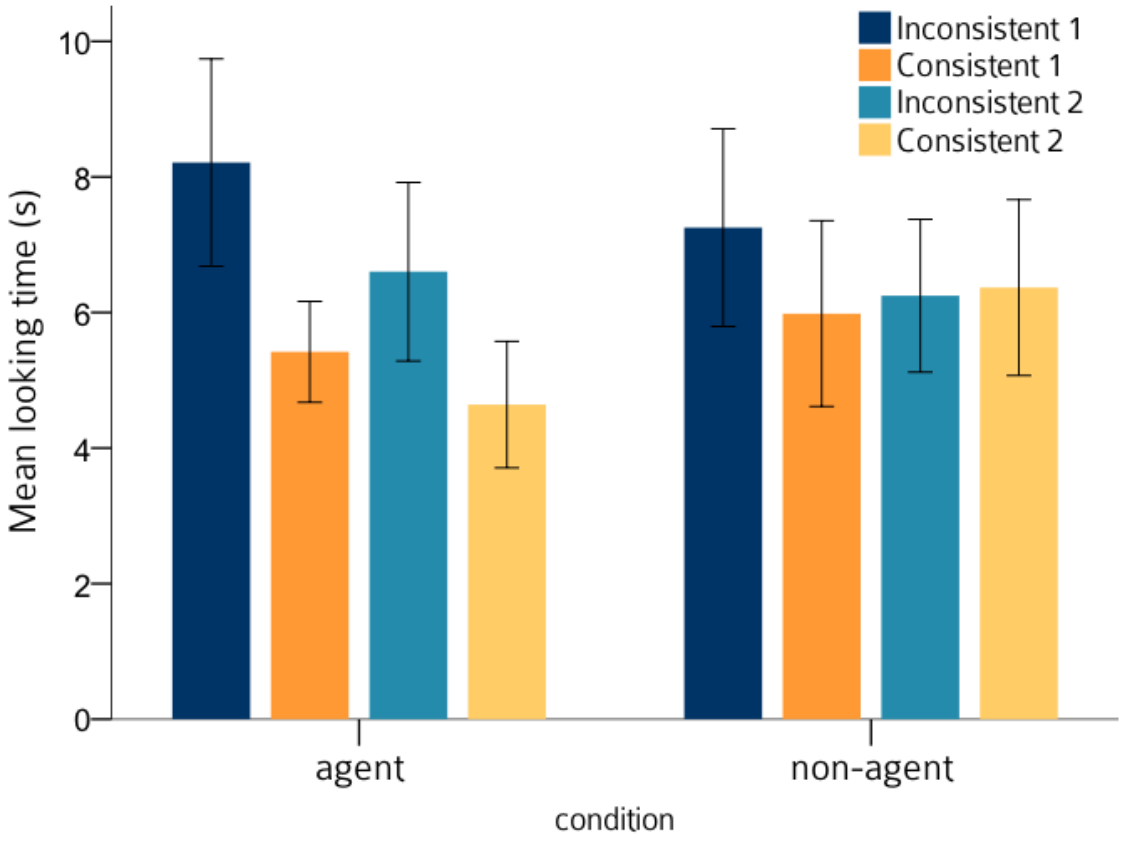

Figure 5.4. Looking times in Experiment 7. Error bars show +/- 1 Standard Error.

We analyzed looking data in a 4-way mixed ANOVA with Trial pair (1 vs. 2) and Choice (IC vs. C) as within-subject variables and Condition (Agent vs. Non-agent) and Order of Test (C first vs. IC first) as between-subjects factors (see Figure 5.4). Should infants keep their attribution of goal-directedness in the Agent condition, but discard it in the Non-agent condition, this would predict a three-way Trial pair*Condition*Choice interaction; as only on Trial pair 2 in the Agent condition should infants' looking times not differ from each other.

Analyses showed that there was an overall main effect of Choice ($F(1,44)= 5.618$, $p= .022$, partial $\eta^2=.113$), with longer times during Inconsistent events ($M_{IC}=7.079$, $SD= 5.272$) than in Consistent events ($M_C=5.603$, $SD= 4.295$). There was also a Choice*Order of Test interaction ($F(1,44)= 4.319$, $p= .044$, partial $\eta^2=.089$). This was due to the fact the effect of Choice was significant in the IC first group ($F(1,22)= 9.593$, $p= .005$, partial $\eta^2=.304$), but not in the C first group ($F(1,22)= .044$, $p= .836$, partial $\eta^2=.002$).

With planned comparisons we assessed whether looking patterns on the first and second test trial pair differed between the two conditions. On each test trial pair we conducted a 3-way mixed ANOVA with Choice (IC vs. C) as within-subject variable and Condition (Agent vs. Non-agent) and Order of Test (C first vs. IC first) as between-subjects factors. On the first test trial pair there was a significant effect of Choice ($F(1,44)= 4.134$, $p= .046$, partial $\eta^2=.087$), with longer times during Inconsistent events ($M_{IC}=7.732$, $SD= 7.316$) than in Consistent events ($M_C=5.702$, $SD= 5.397$). There was no other main effect or interaction. On the second test trial pair, however, there was no main effect of Choice ($F(1,44)= 0.989$, $p= .325$), and also no other main effect or interaction.

As our predictions with regard to differences in infants' looking patterns mainly concerned the second test phase, we analyzed looking behavior on the second trial pair in the two conditions separately. In the Agent condition infants looked longer during Inconsistent choice events (M_{IC}=6.602, *SD*= 6.717) than in Consistent choice events (M_C=4.641, *SD*= 4.571), however, this difference was not significant ($t(23)$=1.402, p= .174). In the Non-agent condition looking times were approximately the same duration in the two conditions (M_{IC}=6.248, *SD*= 5.526; M_C=6.368, *SD*= 6.359), with no significant difference between the two ($t(23)$=-.100, p= .921).

We then log-transformed the looking time data (Csibra, Hernik, Mascaro, Tatone, & Lengyel, 2016) performed on these data the same analyses as before. These analyses yielded similar results as the ones reported on the raw data, however, significance values differed. The main effect of Choice showed a tendency but did not reach significance ($F(1,44)$= 3.649, p= .063, partial η^2=.077), with longer times during Inconsistent events than in Consistent events. The Choice*Order of Test interaction reported earlier was not significant on the log-transformed data ($F(1,44)$= 1.124, p= .161, partial η^2=.044). When analyzed on the two trial pairs separately, the effect of choice did not reach significance on either trial pairs (first test trial pair: $F(1,44)$= 1.124, p= .295; second test trial pair: $F(1,44)$= 2.181, p= .147). Finally, on the second trial pair differences between the two types of test trials (Inconsistent vs. Consistent) were similar as in the raw data (Agent condition: $t(23)$=1.599, p= .124; Non-agent condition: $t(23)$=0.443, p= .662).

5.3.3. Discussion

The present study had two main purposes. First, to assess whether infants perceive an animated object's self-propelled motion as goal-directed. We implemented changes on Experiment 6 to make the stimuli more clear, and the design suitable to test this question. Specifically, we added a test phase after familiarization to see whether we can replicate the Woodward-effect with our stimuli and design. Second, in case infants initially interpret the agent's behavior as goal-directed, we aimed to test whether they can (i) maintain the goal attribution if they face new information that however should not change this interpretation (Agent condition), or (ii) discard this attribution of goal-directedness if they receive evidence suggesting they should change their interpretation of previous events (Non-agent condition).

In the first test phase infants looked longer during Inconsistent choice compared to Consistent choice events. This shows that in this initial phase infants encoded the rectangle's actions as goal-directed, and expected it to continue to act based on its goals, suggesting a successful replication of the Woodward-effect with 3D animations. This provides supporting evidence that self-propelledness and varied behavior (i.e. that the object can change its direction) can elicit the attribution of goal-directed actions.

Infants' looking behavior during the second, critical test phase shows a less clear pattern. When averaged across the two conditions infants did not look longer at the Inconsistent choice events, which would be consistent with both the lack of effect in one of the groups but not the other; or the lack of effect in both. However, there was

no interaction with condition – which would be predicted however, if infants' looking patterns would systematically differ on the two conditions. Moreover, while infants in the Agency condition on the second trial pairs showed a numerical difference in the predicted direction, this did not reach significance. Therefore it is unclear how the absence of effect in the Non-agent condition should be interpreted. This absence of effect would be predicted in case of revision of agency, but is only meaningful in comparison with the *presence* of an effect in the Agent condition. On one hand, it could be that in the Non-agent condition infants indeed revised their agency attribution after observing the second orientation event; and the lack of significant effect in the Agent condition is due to the fact that infants by then have received one test phase and their attention overall dropped.

However, it is also possible that in both conditions the effect simply became weaker by the second test phase. This leaves open the possibility that infants in fact do *not* revise their agency attributions, at least not more when presented with evidence that otherwise does not elicit attribution of goal-directedness (such as the long handle condition in the study of Luo & Baillargeon, 2005). On the other hand, it can also mean that infants, whenever presented with new information, struggle to integrate it into their already existing agent-representation. This, however, is somewhat doubtful, as infants likely encounter such events on a daily basis, when they only partially see an object, or a person, and only later do they see the full appearance. Another alternative is that even the short handle causes infants revise their attribution. In order to exclude the possibility that a handle of any kind disrupted infants' agency attribution, follow-up studies could show the agent with the short handle from the beginning. Should infants attribute goal-directedness in this case, then the lack of effect in the present study is likely due to memory or other

external limitations. However, if infants would not interpret the actions as goal-directed if they observe such an agent with a short handle from the beginning; then the absence of effect would in fact be explained better by a successful revision of agency attribution in *both* of our conditions. In the latter case subsequent studies should establish a feature (that fulfills the role of the handle as a possibly later revealed part of the agent) that when fully visible from the beginning, elicits attribution of goal-directedness. Once established, this feature could be used in a condition where after a second orientation phase the feature would be revealed, and this time goal-attribution should remain.

Relatedly, it could be that the inconsistent choice trial in the first test phase conveyed to infants that the agent's preference for one object over the other is not exclusive. While this is unlikely based on other studies using several repeated test pairs (e.g. Woodward, 1998), in our case there was an additional delay between the test phases which may influence infants' interpretation. In order to avoid such a possibility, and to aid infants' memory and keep them attentive throughout; a next study could leave out the first test phase, and revert back to the trial structure of Experiment 6, but keep the other characteristics of stimuli and design of Experiment 7. This would allow to more directly test the effect of receiving new information about the agent, on infants' attribution of goal-directedness.

Finally, it is possible that it is ambiguous to the infants, whether the manipulation (i.e. showing the short or long handle) reflects the entity's state, or trait. In other words: showing that it can be possibly moved by the handle, does not mean it could not move by itself. Relatedly, the original findings of Luo & Baillargeon (Luo & Baillargeon, 2005) can be interpreted as a sign of self propelledness, as the authors

originally argued; or according to a somewhat different explanation where the handle signals a restricted movement of the box, suggesting it may not have control over its actions (Baillargeon, Scott, & Bian, 2016). We chose to show in test phase the closed occluders to match the events between conditions; however, a strong test would be to show occluders open during test phase; and leaving the handle visible during the approach of the (old or new) target. If infants in this case would not modify their expectations, it would signal a strong endurance of attributions, resilient to counterevidence.

In sum, the present study aimed to investigate whether (i) infants perceive a 3D animated rectangle as a goal-directed agent; and then (ii) when presented with additional information about this agent, they selectively keep or discard the attribution of goal-directed agency, depending on the type of information. We succeeded to show (i) but not (ii), as results are unclear as to what may cause the pattern of infants' behavior. Further studies are therefore needed to disentangle the various factors that are at play when infants observe such events.

5.4. General discussion

The present set of experiments targeted the question whether infants' representations of goal-directed agents are (i) *flexible* enough to be modified or discarded if needed; and whether their representations are (ii) *strong* enough to be maintained if new information needs to be added. Our results so far are inconclusive, as assessing above point (i) would have relied on infants' success on (ii). In the present studies infants, while initially seemed to ascribe goal-directed agency; even

when later presented with confirmatory evidence, did not give evidence of strong expectations of the agent's behavior.

While the empirical assessment is subject to methodological challenges, more broadly the current chapter aimed to contribute to the characterization of the processes involved in mental state representation. Here we addressed this question through goal attribution in infants; however, how information is integrated with existing mental state representations is unexplored with regard to other mental states, as well as other age groups. There may be differences between various mental states how resilient they are to counterevidence. For example, demonstrating a goal or preference may enable learning about desired objects, therefore they may not be encoded as bound to a particular agent (Kampis et al., 2013); however, it could nevertheless be highly sensitive to the source in order to evaluate whether it is a reliable informant. There may also be a developmental trajectory where due to the interplay between other cognitive abilities, such as memory abilities, executive function, or metacognition, the process of mental state calculations may also become more flexible.

The rationale behind the experiments in this chapter is that infants evaluate the representations they formed of an agent, in light of new information. However, a piece of information to someone may only cause the revision of their attribution if they have access to what was their original basis of attribution. In other words, infants in the present studies may only consider the evidence for non-self-propelledness as a reason to revise their attribution of agency, if they know that the basis of their original attribution was self-propelledness of the agent. This may be a challenging process for the young mind, as it might require metacognitive reflection,

and possibly episodic recollection of the basis of their initial inferences. Explicit metacognitive abilities were found to be present at 20 months (Goupil, Romand-monnier, & Kouider, 2016), and even 12-and 16-month-old infants can selectively guide their behavior in order to acquire new information (Begus & Southgate, 2012; Kovács, Tauzin, Téglás, Gergely, & Csibra, 2014). Relatedly, while evidence for explicit memory has been found already at 6 months of age (Barr, Dowden, & Hayne, 1996) episodic memory has been argued to emerge in its rudimentary form around children's second birthday in humans (Newcombe, Lloyd, & Ratliff, 2007). While some argue that metacognition and episodic memory are present in non-human animals (Griffiths, Dickinson, & Clayton, 1999; Kornell, 2009), it is unclear whether infants in their first year of life possess such abilities. Therefore the flexibility of infants' mental state attributions and their relation with other cognitive abilities may be subject of future research.

Chapter 6

General discussion and conclusions

"When someone points out all this mindreading to you, it hits you with some force. […] We mindread all the time, effortlessly, automatically, and mostly unconsciously. That is, we are often not even aware we are doing it – until we stop to examine the words and concepts that we are using."

- Simon Baron-Cohen (1997)
Mindblindness: An essay on autism and theory of mind

Human sociality has been of interest for millennia. Recently the focus of attention has become whether primates are capable of thinking about other's mental states (Call & Tomasello, 2008; Martin & Santos, 2016; Premack & Woodruff, 1978), which inspired psychologists and philosophers to investigate the origin of the ability termed 'Theory of Mind' in human ontogeny (Baillargeon et al., 2010; Caron, 2009; Leslie, 2000a; Wellman et al., 2001; Wimmer & Perner, 1983). Research from the past decade has shown that human infants spontaneously represent others' mental states, (Kovács et al., 2010; Onishi & Baillargeon, 2005; Southgate & Vernetti, 2014) which is in apparent contrast with initial findings suggesting a later emergence of ToM in preschool age (Perner et al., 1987; Wellman et al., 2001), and with studies showing that adults in some cases don't encode others' beliefs or perspective (Keysar, Lin, & Barr, 2003; Low & Watts, 2013; Surtees, Samson, et al., 2016). To account for these differences, some theories propose that infants (and possibly nonhuman animals) can behave in a way that approximate a metarepresentational understanding of other minds, but the underlying cognitive systems lack some of the core characteristics of fully elaborate ToM (Apperly & Butterfill, 2009; Butterfill & Apperly, 2013; Perner et al., 1987, 2011; Rakoczy, 2012). Similar mechanisms were suggested to operate when people engage in fast, spontaneous, or implicit computations of mental states (e.g. Samson, Apperly, Braithwaite, Andrews, & Bodley Scott, 2010).

In contrast, several theorists have argued that core mindreading abilities are present from birth, and continue to operate into adulthood (Carruthers, 2013; Kovács et al., 2010; Leslie, 1987, 2000a). Most of these proposals have challenged the necessity to posit different concepts to describe infants' ToM abilities, and suggest a

view of mindreading as a synchronized functioning of several sub-processes, some of which may be domain-specific to ToM, and other domains or more domain-general processes may work in concert with them (Carruthers, 2015b; Christensen & Michael, 2016; Kampis et al., in press, 2015; Scholl & Leslie, 1999). This multi-componential view is echoed in the construct of belief files (Kovács, 2015), that aim to provide an account how elements of belief representations may be handled separately, in a fast and efficient manner. Belief files, combined with other file-like structures that were proposed to capture mental representations (Kahneman et al., 1992; Perner & Leahy, 2016; Recanati, 2012; Scholl & Pylyshyn, 1999), may provide a framework to the multi-componential functioning of ToM.

In light of these considerations, the present work aimed to take a step towards the characterization of the mental representations and cognitive processes underlying Theory of Mind in infants. It explored the possibility that ToM abilities in their core format are present from early on, and asked: what cognitive mechanisms may enable representing mental states infants? On one hand, it proposed possible mechanisms through which belief contents are handled, and provided supporting electrophysiological and behavioral evidence. On the other hand, it outlined uncharted aspects of mindreading with regard to the flexibility of the underlying processes. Lastly, it provided empirical investigation of predictions from alternative theories claiming conceptual change in ToM abilities. Specifically, the work in this thesis explored the following questions:

- Do signatures of representing primary representations also accompany representing them as belief contents?

- Does the modulation of one's own behavior by others' mental contents manifest itself in infants' active behavior?
- Can infants represent beliefs with a variety of contents that they themselves can entertain? Do infants show arbitrary limits in their ToM abilities, or do limitations correspond to their own representational abilities?
- Do infants represent mental states on-line, and can they flexibly modify them in light of new information?

This chapter summarizes the experimental findings presented in this thesis and discusses them in connection with each other and theoretical implications, as well as outlines potential limitations and topics for future research.

6.1 Summary of findings

The present thesis started with proposing a possible mechanism through which the contents of mental states may be metarepresented in the infant mind. Based on the idea of re-using primary representations as the contents of metarepresentations (Carruthers, 2013; Leslie, 1987; Wilson, 2000), we have put forward the proposal that infants recruit their cognitive mechanisms dedicated to represent the environment to represent others' mental state contents. This notion is supported by findings showing that encoding particular events from infants' own and from someone else's perspective recruit analogous cognitive processes (Rueschemeyer et al., 2015; Southgate & Vernetti, 2014).

We argued that shared underlying mechanisms for one's own and attributed representations predict that signatures of representing primary representations should also accompany representing them as belief contents. Measuring such signatures may provide a useful tool to investigate in a wide range of scenarios whether an observer metarepresents another person's mental content. One current question is whether in perspective taking people in fact attribute representations to others. We suggested that the modulation effect, whereby a person's behavior is influenced by another person's visual perspective or belief (Kovács et al., 2010; Samson et al., 2010), is best explained by representing the other person's representation, which then modulates one's own behavior (due to their shared representational format). To test whether in perspective taking infants represent another person's object representation, we built on earlier paradigms that found a specific brain signature accompanying object representations in infants; gamma-band oscillatory activation in electroencephalographic responses over the temporal regions (Kaufman et al., 2003, 2005). We observed such gamma-band activity in 8-month-old infants' EEG when an object was occluded from the infants' perspective, as well as when it was occluded only from another person (Experiment 1), and when subsequently the object disappeared but the person falsely believed the object to be present (Experiment 2). This suggests that infants attributed a sustained representation of the object to the actor on-line, when she lost visual access to the object, and further sustained this representation when the other's belief became false. In addition, these results give support to the hypothesis that infants utilize their own representational system that is otherwise used for encoding objects and events in the world, to ascribe representations to others.

Building on these findings, we hypothesized that the range of belief-involving scenarios infants can handle is much wider than what has been investigated so far. Specifically, we proposed that any particular infant should be able ascribe to others representations that she herself can entertain. Experiments 3-5 investigated this by exploiting the rich body of findings on infants' knowledge on the physical world of objects. Infants possess powerful cognitive mechanisms that enable representing and tracking objects around them in relatively complex scenarios. They can individuate and track multiple objects at a time, and through their first year of life they can store more and more information about these objects (Leslie et al., 1998; Wilcox, 1999).

To start exploring whether infants utilize these abilities in social settings, Experiment 3 adapted a manual search task that was used by Feigenson & Carey (2003). In their studies 14-month-old infants were found to successfully track objects that were hidden in an opaque box up to the limit of 3 items, and perform simple calculations with them. This was evident in longer search duration when an object still remained in the box (e.g. 2 were hidden but only 1 retrieved), compared to when all objects were retrieved. In Experiment 3 we asked whether infants at this age can also track such scenarios from someone else's perspective, and successfully represent if another person has a false belief about the content of the box. To measure this, we turned to the modulation effect and hypothesized that infants' search durations, similarly to other continuous measures (van der Wel et al., 2014), may be modulated by the other person's belief. We created scenario pairs that were matched with regard to the infants' knowledge about the content of the box (i.e. always empty or always containing one object), but manipulated what the other person believes to be inside; and compared infants' behavior within the scenario pairs. Results from infants' manual search duration showed that infants indeed searched longer in the

box when the other person falsely believed an object to be present, compared to when she knew that all objects were retrieved. Moreover, infants seemed to search less in the box, when there was in fact an object inside, but the other person false believed that all objects were retrieved, compared to when she knew there was an object left to obtain. This latter finding raises the possibility that infants represented the other person's belief about absence (as simply not imputing any representation to the other would not modulate infants' behavior), but without appropriate controls this remains a question that may be subject of further investigation. Together, findings from Experiment 3 suggested that infants successfully individuated and tracked one or two objects based on spatiotemporal information, and were able to use these inferences to represent another person's false belief involving these objects.

This led to the question whether infants can take into account not just spatiotemporal but feature/kind information to individuate and track objects form someone else's perspective. Earlier studies involving the same manual search paradigm from Feigenson and Carey (2003) found that from their own perspective, when spatiotemporal information is ambiguous, infants can rely on feature/kind information (Zosh & Feigenson, 2012). Moreover, they can vary whether to rely on such cues based on their previous experience whether two different appearances indeed indicate two different objects (Cacchione et al., 2013). Can infants operate with the same inferences when computing someone else's belief? Experiment 4 explored whether 14-month-old infants can successfully infer that someone can have a false belief based on a correct individuation of two objects that are of different appearance. Building on this, Experiment 5 asked whether infants can ascribe to someone a false belief based on that person's mistaken individuation of two appearances of one object into two objects. Similarly to Experiment 3, we found that

infants' search times in Experiment 4 and 5 were modulated by the other person's belief. Crucially, in Experiment 5 this suggests that infants represented the other person's mistaken belief about object identity; which speaks against recent proposals not granting infants such abilities (Butterfill & Apperly, 2013).

The fact that infants can represent others' beliefs with a variety of previously unexplored contents suggests a relative flexibility of the processes involved. In addition, Chapter 2 showed that infants manipulate the represented belief contents on-line, as events are unfolding. However, while those results imply that infants formed and sustained an attributed representation at the time of the relevant changes, there may be also cases when new information needs to be added to the already existing representations. Such information may simply need to be integrated, but occasionally it may also invalidate the representation. This led to the question how flexible ToM processes are with regard to radical revision or discarding already computed representations, which motivated Experiments 6-7 in Chapter 5.

While much research on ToM is concerned with belief understanding, many of these studies in fact rely on the assumption that infants understand goals, preferences, or desires. When infants' expectation of an actor's actions is assessed, this expectation is often based on the agent's goal to get (or a preference for-) a specific target (e.g. Onishi & Baillargeon, 2005). Therefore assessing whether infants can update or discard belief representations has to follow testing first whether they can perform these operations on goal or preference attributions. Otherwise if infants were found not to expect an agent to act towards a goal based on her belief, this could be due to failure of updating the belief representation, the goal representation, or both. Experiments 6-7 aimed to contribute to characterizing the flexibility of mental

state representations, by asking whether infants can revoke their existing goal ascriptions when presented with evidence suggesting that their original attributions were mistaken. We created a scenario based on a frequently used paradigm to measure infants' expectation of goal-directed action (Woodward, 1998). Using a modified version of the Woodward-paradigm Luo and Baillargeon (2005) found that 5-month-old infants interpret the actions of a self-propelled box as goal-directed, but only if the box's appearance was consistent with it moving by itself. If the box had a handle attached to it that reached outside the screen, infants did not read its action as goal-directed.

We adapted this paradigm and presented infants initially with a scenario that matched the above condition where we can expect goal attribution to a seemingly self-propelled rectangle-shaped box, only to then reveal to them later that they have missed some information so far. This information was either consistent or inconsistent with the box moving by itself; therefore obtaining this new information should result in keeping the agency and goal attribution in one case but not the other. Presenting the events this way allowed us to ask whether infants' representations of goal-directed agents are (i) *flexible* enough to be modified or discarded if needed; and whether their representations are (ii) *strong* enough to be maintained if new information needs to be added. However, assessing above point (i) relies on infants' success on (ii), therefore the inconclusive results on infants' ability to maintain representations when new information should be integrated, make it challenging to interpret findings regarding the flexibility of representations. In the present studies infants, while initially seemed to ascribe goal-directed agency; even when presented with confirmatory evidence, did not give evidence of strong expectations of the agent's behavior. The potential reasons behind this pattern of findings are discussed,

such as general attention effects, interference from multiple points of assessment, infants' metacognitive or episodic memory limitations, or whether the information to be incorporated confirms goal-directedness may be ambiguous to infants. While the empirical assessment is subject to methodological challenges, Experiments 6-7 aimed to contribute to the characterization of the processes involved in mental state attribution.

In sum, the present thesis investigated with various methodologies the cognitive processes involved in representing others' mental states in infants. First, it showed that signatures of primary representations also accompany representing them as belief contents (Chapter 2). Second, it explored the modulation effect where ascribed representations influence one's own behavior, and demonstrated that infants can handle belief representations with a variety of contents that they themselves can entertain (Chapter 3 & 4). Last, it probed the flexibility of ToM processes when updating or revision of existing representations is necessary (Chapter 5).

6.2. Common format of own and attributed representations

Theoretical implications

We proposed in the introduction of this thesis that a cost-effective way to handle metarepresentations of attributed mental states would be to re-use one's primary representations as the content of these representations. This is in line with findings in Chapter 2, showing shared signatures for own and attributed representations; and Chapter 3 & 4, showing a modulation of infants' own behavior by the other person's belief. We outlined a model whereby multiple levels of file-like structures would constitute a belief representation or 'belief file' (Kovács, 2015), which could take as

content mental files (Perner et al., 2015). Mental files could contain various kinds of information, such as representations of objects; in which case they may be linked to object files (Kahneman et al., 1992), which in turn are indexed to external objects.

In what format, then, did infants represent their own representations, and the ones they ascribed to the other person? The findings that provided the basis for our studies (Feigenson & Carey, 2003; Kaufman et al., 2005) argued that infants represented the objects in such scenarios via object files. However, object files most likely cannot be ascribed to others, as (i) they are mid-level representations in the visual system, and (ii) two indexes (e.g. one's own index and an 'ascribed index' linked to the same object) cannot occupy the same point in space. Therefore such representations likely need to be ascribed via another mechanism, for instance as outlined earlier, via (vicarious) mental files attributed to another person. However, if infants' own representation of an object is an object file, and the one ascribed to the other person is a (vicarious) mental file, then how do these relate to each other? It is possible that once a vicarious mental file is formed, it prompts one's own representation to also be represented as a mental file. Alternatively, vicarious mental files and object files may be similar enough to account for our results. For instance, vicarious mental files could be referenced to the object files (object files integrated into mental files), or one can even think of a cognitive system that can deal with both. After all, vicarious mental files are simply attributed mental files, and mental files are constructs that are suggested to guide infants' behavior. Therefore at any given moment infants' own and the ascribed representation do not have to be represented on the same level (object file vs. mental file), and hence the modulation effect might be due to vicarious files resembling be infants' own representations.

Relatedly, the activation patterns in Chapter 2 could be due to the fact that a mental file that is ascribed to the other person is in fact attached to infants' object file (the access of which was shown by Kaufman *et al*., 2005, to elicit the activation pattern we also obtained). This raises the intriguing question how others' representations are sustained once one's own representation is discarded. In Experiment 2, the object disintegrated, visibly to the infants; which should cause infants to discard their representation if it was an object file (Kaufman et al., 2005; Scholl & Pylyshyn, 1999). However, our results suggest that infants showed the same activation as when they themselves sustain the representation of an object. Since we argued that object files *per se* couldn't be ascribed to others, this activation has to signal the sustaining of the others' vicarious mental file. Therefore our results point to the possibility that gamma-band oscillations could reflect the sustaining of mental files as well. Clarifying the role of the cognitive systems involved in such computations therefore may be the subject of future work.

Differentiation of own and attributed representations

The findings that one's own representations are represented in a shared format as attributed representations raise the question how infant's primary representations would be separated from the representations ascribed to others. While the present studies do not directly address this question, in Study 1 & 2 in addition to our predicted activation patterns when the presented object became occluded from the infants' or the other person's view; we observed an additional activation that accompanied only processing the object occlusion from the actor's perspective, in both studies. The fact that similar activation did not occur during the Occlusion from the Infant events suggests that it might reflect some further processing of ascribed

representations, and could potentially play a role in distinguishing an ascribed representation from the infants' own reality representation.

Primary and attributed representations may differ in many respects. One aspect that set them apart is their distinct functional roles in cognition. One's own representations serve the function of representing one's knowledge of a situation, subserve learning and making predictions about the environment, and may feed into various inferential processes, and infants' own motor planning. Attributed representations enable the prediction of others' actions, and participating in communicative acts with others; therefore they may feed into other ToM components, or others' action prediction. For instance, seeing an agent may trigger representing its mental state (e.g. set up a belief file) (Leslie et al., 2005). This requires calculating the content of the belief; which may be done by the various cognitive systems that subserve representing that content (e.g. if an object is hidden, this would involve the object tracking system, like in Chapter 2). Once this content is represented, it can feed into action prediction, such as to predict whether the agent will reach for the object (Southgate & Vernetti, 2014). Relatedly, a proposed subsequent step after belief formation is a selection of the correct representation (Leslie et al., 2005; Moll, Meltzoff, Merzsch, & Tomasello, 2013; Qureshi, Apperly, & Samson, 2010). This is in line with electrophysiological findings showing that in a perspective taking task (based on Samson et al., 2010) there is a late slow wave over right frontal areas that is sensitive to the inconsistency between one's own and the other's perspective (McCleery, Surtees, Graham, Richards, & Apperly, 2011). Others have found a late slow wave over frontal areas that is sensitive to judgments of belief vs. reality, albeit lateralized to the left hemisphere (Liu, Sabbagh, Gehring, & Wellman, 2004).

Further studies using EEG may look at the temporal patterns in different brain regions that contribute to the processing of belief scenarios, to clarify the functional roles of certain systems. For instance, our results in Chapter 2 do not clarify whether the gamma-band activation has a functional role in sustaining the representation; or rather indicates that infants are accessing or manipulating the attributed representation. However, recent findings with 6-8-month-old infants suggest that maintaining the representation of two objects results in greater gamma-band activity than maintaining one (Leung et al., 2016), albeit in more occipital regions compared to the findings observed in the temporal regions, reported in Chapter 2; or in earlier studies investigating object occlusion (Kaufman et al., 2005). Since gamma oscillations increase with working memory load (i.e. number of objects held in working memory), this could provide a tool for future investigations on the relationship between working memory and belief representations. For example through testing whether if someone represents another person's perspective (in addition to representing it from one's own perspective), gamma power increases as it does when one represents two objects from their own perspective, it could be assessed whether own and attributed representations draw on similar working memory resources.

6.3. The modulation effect

Theoretical implications

We argued in the introduction that the similar format of primary and attributed representations explains the modulation effect. In three studies we indeed showed this effect, whereby infants' behavior in a manual search task was influenced by another person's false belief. These findings may seem counterintuitive. Why do infants behave based on a representation that they ascribe to another person? One possibility would be that infants simply cannot bind the belief contents to the respective person. However, behavioral (Southgate et al., 2007) as well as electrophysiological (Southgate & Vernetti, 2014) evidence suggests that if infants observe an agent, they are able to correctly predict her actions based on her beliefs. A possible model would be that different components of the belief representations are linked together, and inferences 'travel' through these links. The main elements of these representations, as discussed in the introduction, could be an agent, a belief content, a relation between these two, and an external anchor that links the belief content to the environment (i.e. a referent). In case of action prediction the chain of inferences may start from the agent (whose action one wants to predict, say, John); then go through the belief (e.g. he believes there is an object in the box) and then the anchor (an index to the box), in which case it will lead to successful action prediction (John will reach into the box). However, it may be different in the case where an object, i.e. a referent/anchor, is visible at the moment of inference. In such a case all representations linked to that object may be activated; for instance one's own, and any attributed belief content. These contents would need to be handled differently based on which agent they belong to; such as selecting one's own representation to formulate one's own motor plan. However, this selection could be a subsequent,

resource-demanding step (cf. Leslie's SP, or selection processor; whose task is to select among possible representations), which might not always be executed (due to lack of resource or motivation). Therefore the modulation effect is unlikely to be due to infants' (or adults') inability to separate one's own and attributed representations, but rather it arises in situations when both representations become activated and selection between them does not yet take place.

In addition, the modulation effect may be related to the development of infants' memory abilities. It seems plausible that encoding and successfully retrieving the agent to whom a representation 'belongs to' may involve processes related to episodic memory; the common component being the necessity to connect various aspects of a situation to each other. Indeed, it has been suggested that infants in the first two years have extremely limited episodic memory capacities (Newcombe et al., 2007). However, as the modulation effect has been observed in adults as well, developmental factors cannot fully account for the modulation effect. It may be that spontaneous mindreading elicits less retrieval of episodic properties, therefore the connection between the representation and its *source* (i.e., the respective agent) may be less strong. In fact, the source monitoring approach in memory research proposes that people "do not typically directly retrieve an abstract tag or label that specifies a memory's source, rather, activated memory records are evaluated and attributed to particular sources through decision processes performed during remembering" (Johnson, Hashtroudi, & Lindsay, 1993, p. 3). One aspect of source monitoring is "external source monitoring", that entails for instance distinguishing memories that are based on what person A stated, from what person B stated. While this framework is mostly concerned with explicit judgments on the origin of specific reportable memories; they emphasize that these decisions rely on different characteristics of

memories from various sources; and should these characteristics not be encoded or retrieved, it would result in deficits of source monitoring (Johnson et al., 1993). Therefore the separation of stored representations in explicit reports may depend on whether the necessary information is retrieved. If such information is lacking, it may result in some situations in a less pronounced distinction and hence increased penetrability between representations – such as between one's primary representations and attributed ones. This, while a far analogy; could potentially account for developmental patterns (e.g. the lack of binding mental states to agents as suggested in Kampis, Somogyi, Itakura, & Király, 2013) as well as the modulation effect in implicit/spontaneous ToM tasks. The fundamental distinction between implicit or spontaneous mindreading and elicited or explicit mindreading may be the amount of characteristics retrieved about a representation. This is in line with the suggestion of Elekes *et al.* (2016), who propose that Level-2 perspective taking may not always happen spontaneously because it would involve the others' representation of object *features*; and encoding such information may not be triggered simply by cues indicating that the other is aware of the *presence* of an object, but may depend on receiving information that the other is attending to the aspectual properties of the object. Spontaneous belief attribution may therefore occasionally retrieve less detailed information with regard to how the representations (memories) are formed (or other characteristics), and therefore may not manifest itself in Level-2 perspective taking / aspectuality ToM scenarios (Low & Watts, 2013; Surtees, Butterfill, & Apperly, 2012; Surtees, Samson, et al., 2016); or may not be reportable (Schneider, Bayliss, Becker, & Dux, 2011).

Relationship of the modulation effect with other capacities

To understand the modulation effect better, it may thus provide insights to investigate how it relates to overall ToM abilities, and to other cognitive capacities. On one hand, it could be that infants who have better overall ToM capacities show a (larger) effect, and the ones who don't show the effect are ones who simply don't compute the other's mental state in these cases. On the other hand, it is possible that all infants represent the others' belief, but better inhibitory capacities enable better suppression of the modulation, therefore infants who show the effect and ones who don't, differ in some other capacity than ToM (e.g. executive function). Finally, it could be that whether infants successfully represent the other's belief, depends on capacities outside of ToM, such as better working memory resources enabling the sustaining of such beliefs. The present study is not suitable to distinguish between these alternatives. However, a future possibility is to compare infants' behavior in belief-involving settings with their performance on other tasks, either at the same age when they take part on the false belief task, or to follow up later to assess stability in individual variations and connection with other capacities. In addition, such swift availability of the other's mental state may enable the acquisition of socially relevant information. If so, then the way infants perceive the agent (e.g. knowledgeable vs. unreliable) may influence the occurrence of the modulation effect. In line with this, infants follow selectively the gaze of a reliable social partner when she looked behind a barrier, compared to a previously unreliable one (Chow, Poulin-Dubois, & Lewis, 2008); and they are less likely to attribute a belief to a person whose gaze previously deemed unreliable (Poulin-Dubois & Chow, 2009). Therefore infants may selectively vary the readiness of attributing beliefs to reliable and unreliable social partners; but

possibly even if they compute an unreliable partner's belief they may be more likely to separate their beliefs from their own.

6.4. Limits of early ToM

Registrations and mental files

Several findings reported in this thesis speak to the proposed limitations of early mental state reasoning. First, according to the minimal-ToM theory (Butterfill & Apperly, 2013), instead of beliefs 'proper' infants represent registrations; which are part of a fast and efficient, but limited mindreading system, or 'System 1'. Registrations are relations between an agent and an object, which cannot grasp aspectuality (i.e. *how* someone represents an object); therefore infants should not have the ability to represent others' beliefs about objects. Relatedly, System 1 was suggested to operate in spontaneous mindreading even in adults, therefore representing beliefs about identity should not occur spontaneously. In contrast with these proposals we found that infants represented another person's belief about object identity in a spontaneous task.

Second, findings in Chapter 3 tentatively suggest that infants represent others' beliefs about absence of objects. There, even though there remained an object to search for, if the other person falsely believed that all objects were retrieved, infants searched less (compared to when the other person knew there was still an object present). Should these results hold against the appropriate controls (ruling out alternatives such as infants being more prone to search if they *and* someone else believe an object to be present), this may posit challenges to the minimal-ToM view. Since registrations track the location of an object, they may enable *positive* tracking

(e.g. an agent having registered an object in location X), but they could not represent *negative* tracking (such as an agent having registered where the object *is not*); and within this system it does not seem possible to represent the *absence* of a registration, nor to represent beliefs whose content can only be expressed via quantifiers or negation (e.g. there is *no* toy in the box).

Third, neither registrations (Apperly & Butterfill, 2009; Butterfill & Apperly, 2013), nor vicarious mental files that are linked to one's own, regular mental files (Perner et al., 2015; Perner & Leahy, 2016) would allow for attributing representations about non-existent objects. However, in Experiment 2 in the second Segment infants likely discarded their own representation, but sustained the attributed belief content. There, the object was first occluded from the other person, and then – visibly to the infants – it disintegrated. Such disintegration events were previously suggested to result in discarding object files (Kaufman et al., 2005; Scholl & Pylyshyn, 1999). If infants discarded their own object representation, but still sustained the other's, it means that the attributed representations do not rely on primary representations. Since registrations are relations to objects, they cannot handle cases when there is no object to link to. Therefore our results make it unlikely that infants represented the other's perspective via registrations, and suggest that at least in some cases infants can operate with other kinds of representations. Mental files would be suitable to operate in a way that may account for our results. However, for this, vicarious files would have to be independent enough from regular files, to possibly continue to exist even if the regular file is discarded. For instance, they may not need to be linked to the regular file in order to be represented. Alternatively, it may be that just infants' object file was discarded, and not the infants' mental file, but

in this case the content of the mental file would need to take some other format then an object file.

Overall, the results discussed in this section suggest that infants may handle other's mental states via representational formats that can handle contents involving not just location, but identity and absence of objects as well; and represent beliefs about objects that ceased to exist. Together, this calls for a refinement of theories on early ToM and the exploration of further characteristics of infants' mental state representations.

Can infants ascribe mental states involving approximate magnitudes?

In the experiments discussed in the present thesis the number of objects to be tracked never exceeded the limit of the object tracking system (Feigenson & Carey, 2003; Scholl et al., 2001). There is, however, another suggested cognitive system that is responsible for representing large, approximate numerical magnitudes (Feigenson, Dehaene, & Spelke, 2004); which likely enables infants to represent abstract numbers from birth (Izard, Sann, Spelke, & Streri, 2009). This system can allow computations that cannot be solved through tracking individual objects, and can handle relations between two large numerosities. Infants can discriminate between large number sets, and the necessary ratio between sets is developing, with 10-month-old infants succeeding at the discrimination of the ratio of 2:3 (Xu & Arriaga, 2007; Xu & Spelke, 2000). Moreover, at least by 9 month they were found to do addition on large sets (McCrink & Wynn, 2004), by 6 month they can abstract ratio relationships between arrays (McCrink & Wynn, 2007); and when given a choice to select one of two large

arrays, they show a preference for the larger set (vanMarle, 2013). But could infants attribute to others mental states about large quantities? Such attributions could not be represented via registrations, as there cannot be a relation between an agent and 'many' objects, or 'more' objects, therefore it would involve the use of quantifiers. For instance, since infants have a preference for the larger amount of objects from their own perspective, it is reasonable to assume that they can attribute to others similar preferences. This could be tested via a Woodward-like paradigm (Woodward, 1998), which was also the basis of the experiments in Chapter 5. Initially infants could initially witness an actor observing two large arrays (between trials of varying exact quantities, but all above 3) disappearing in two different containers, and then choosing the container with the larger array (i.e. containing 'more' items). This should induce in infants attributing to her a preference for more items; which then could be assessed with showing infants a scenario where she would choose the smaller array (inconsistent choice trials) or again the larger one (consistent trials). If infants successfully extracted the information that the person prefers the larger quantities, they should be surprised if she then chooses the container with the smaller array. This, in turn, would suggest that unlike suggested by the minimal-ToM account (Apperly & Butterfill, 2009), infants are capable of representing beliefs involving quantifiers such as *more* or *most*.

Can infants attribute individuation?

In addition to what kind of representations can serve as contents of mental states that infants can ascribe to others, the Experiments in Chapters 3 & 4 raise the question whether infants can attribute inferences to others. Infants in our studies

attributed a certain representation to the other person (e.g. that she represents two objects in the box). The fact that the person believed two objects to be in the box was a result of an individuation process, whereby based on certain characteristics of the situation she inferred two objects to be present. This information could be spatiotemporal (Experiment 3) or feature/kind information (Experiment 4 & 5), and the individuation could be correct (Experiment 3 & 4) or mistaken (Experiment 5). Infants attributed the output of an individuation inference that was based on one of the above type. But it is unclear whether infants in these cases attributed these inferences to the other person. Infants may have 'simulated' an individuation inference, based on the input from the other person's view (e.g. she saw objects at two distinct spatial locations; or she saw two different appearances and does not possess information on the spatiotemporal characteristics). Note that such simulation, again (as discussed in the Introduction), would be predictive, as infants would use one representation attributed to the other person, and feed it into their own inferential processes to predict the correct output (similarly how they ascribe actions based on beliefs). How much do infants attribute of these inferences to others? To the least, infants are able to engage in different object-individuation inferences for the other person based on her experience, when the input information differs from their primary representations (e.g. in Experiment 5 only they - but not the other person - have additional spatiotemporal information that overwrites the feature/kind information of the two appearances). Whether or not infants are able to attribute inferences that guide individuation (or other inferences based on logical operators) to others, cannot be decided based on our findings, and may be addressed by future work.

6.5. Flexibility of ToM

On-line computations of mental states

It has been suggested that mindreading faces two competing demands: on one hand it needs to be fast enough to guide one's actions in on-line interactions; and on the other hand it needs to be flexible enough to enable representing a wide range of mental contents (Butterfill & Apperly, 2013). We have argued in the introduction that these two demands do not have to be contradictory, and outlined a possible representational structure that in principle can enable fast and efficient computations. This structure in case of belief representations is the belief file (Kovács, 2015), which may enable flexible modification of the various elements (the agent and content variables) the representation of which is subserved by the observer's cognitive processes dedicated to represent the environment. Such a model of belief representations needs to be investigated and spelled out in detail. However, there are several aspects of the findings reported in the present thesis that are in line with this proposal. First, as it predicts, infants did not show limitations specific to proposed mindreading systems but rather limits corresponding to their own representational capacities (Experiment 5). Second, infants seemed to track events on-line, as the events were unfolding, and showed signs of sustaining an attributed representation at the time corresponding to (i) when the other person likely performed the same computation, and (ii) when they showed signs of sustaining the representation from their own perspective (Experiment 1). Third, they readily performed such manipulations on the attributed representation independently of the status of their own representation (Experiment 2). Finally, while calling for further investigation, infants may readily discard or overwrite their representations when

presented with evidence that invalidates sustaining the representation (Experiment 6 & 7). It seems therefore that the proposed framework may be a useful tool to characterize early mindreading abilities.

What are the rules governing mental state computations?

Regardless of the vast amount of literature on Theory of Mind, there seem to be at least as many questions, as there are answers. It has yet to be explored, how various ToM sub-components are integrated with each other. As we have proposed, neuroimaging and electrophysiological investigations may lead us closer by shedding light on the temporal dynamics as well as the various brain areas involved in mental state reasoning. Then, as suggested in this discussion, various other capacities may influence ToM cognition, such as memory processes. Investigating the ontogenetic development of specific types of memory abilities may play a crucial role in understanding both the development of representational abilities as well as task performance. For instance, working memory abilities need to be characterized in relation to mental state representations, as there may be limitations that working memory puts on mindreading. In turn, mental state representations may also play a role in working memory processes, such as chunking of stored information. In addition, episodic and semantic memory characteristics might provide useful insights on mindreading processes, both through external limitations depending on the type of memory processes involved in a mental state computation, as well as through facilitating the development of mindreading processes in human ontogeny.

6.6. Conclusions

"In all things of nature there is something of the marvelous."
- Aristotle

Young infants face a variety of challenges. They have limited motor skills, blurry vision, and a rudimentary set of cognitive abilities. An essential part of their life is that they are surrounded by other people on a daily basis, and they are growing up as part of a culture. In order to manage in the complex web of social interactions and relations, humans need to appreciate that others around them are not simply guided by observable events but also by their psychological states. Adults engage in highly sophisticated forms of reasoning about mental states. But maybe even more impressive is the possibility that young learners may already possess some understanding of others' mental lives. When do infants come to have an understanding of other minds? Are their socio-cognitive capacities fundamentally limited, and go through a radical change in the first few years? While non-human species seem to never acquire the necessary cognitive machinery, much evidence points to the possibility that infants from early on may have the capacity to form metarepresentations of others' mental states. This could enable them to efficiently take part in interactions, acquire languages, and learn about the material and cultural environment that surrounds them. By characterizing some of the underlying mechanisms and highlighting challenges that may guide future work, this thesis hopes to have stimulated further research on how living beings understand other minds.

References

Apperly, I. A. (2010). *Mindreaders*. *Mindreaders: The Cognitive Basis of "Theory of Mind."* http://doi.org/10.4324/9780203833926

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, *116*(4), 953–970. http://doi.org/10.1037/a0016923

Apperly, I. A., Riggs, K. J., Simpson, A., Chiavarino, C., & Samson, D. (2006). Is belief reasoning automatic? *Psychological Science*, *17*(10), 841–844. http://doi.org/10.1111/j.1467-9280.2006.01791.x

Apperly, I. A., & Robinson, E. J. (1998). Children's mental representation of referential relations. *Cognition*, *67*, 287–309. http://doi.org/10.1016/S0010-0277(98)00030-4

Apperly, I. A., & Robinson, E. J. (2001). Children ' s Difficulties Handling Dual Identity. *Journal of Experimental Child Psychology*, *397*(4), 374–397. http://doi.org/10.1006/jecp.2000.2571

Aschersleben, G., Hofer, T., & Jovanovic, B. (2008). The link between infant attention to goal-directed action and later theory of mind abilities. *Developmental Science*, *11*(6), 862–8. http://doi.org/10.1111/j.1467-7687.2008.00736.x

Austin, K., Theakston, A., Lieven, E., & Tomasello, M. (2014). Young Children' s Understanding of Denial. *Developmental Psychology*, *50*(8), 2061–2070.

Baillargeon, R. (1986). Representing the existence and the location of hidden objects: object permanence in 6- and 8-month-old infants. *Cognition*, *23*(1), 21–41.

Baillargeon, R. (2004). Infants ' Physical World. *Current Directions in Psychological Science*, *13*(3), 89–95.

Baillargeon, R., & DeVos, J. (1991). Object permanence in young infants: further evidence. *Child Development*, *62*(6), 1227–46. Baillargeon, R., Scott, R. M., & Bian, L. (2016). Psychological Reasoning in Infancy. *Annual Review of Psychology*, *67*(1), annurev-psych-010213-115033. http://doi.org/10.1146/annurev-psych-010213-115033

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences*, *14*(3), 110–118. http://doi.org/10.1016/j.tics.2009.12.006

Baillargeon, R., Spelke, E. S., & Wasserman, S. (1985). Object permanence in five-month-old infants. *Cognition*, *20*(3), 191–208. Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. MIT Press.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*(1), 37–46.

Baron-Cohen, S., O'Riordan, M., Stone, Valerie E, Jones, R., & Plaisted, K. (1999). Recognition of Faux Pas by Normally Developing Children and Children with Asperger Syndrome or High-Functioning Autism. *Journal of Autism and Developmental Disorders*, *29*(5), 407–418. http://doi.org/10.1023/A

Barr, R., Dowden, A., & Hayne, H. (1996). Developmental Changes in Deferred Imitation by 6-to Z-Month-Old Infants. *Infant Behavior and Development*, *19*, 159–170. http://doi.org/10.1016/S0163-6383(96)90015-6

Barrett, H. C., Broesch, T., Scott, R. M., He, Z., Baillargeon, R., Wu, D., … Laurence, S. (2013). ESM for Early false-belief understanding in traditional non-Western societies. *Proceedings. Biological Sciences / The Royal Society*, *280*(1755), 20122654. http://doi.org/10.1098/rspb.2012.2654

Becker, C. A. (1979). Semantic context effects in visual word recognition: An analysis of semantic strategies. *Memory & Cognition*, *8*(6), 493–512. http://doi.org/10.3758/BF03213769

Begus, K., & Southgate, V. (2012). Infant pointing serves an interrogative function. *Developmental Science*, *15*(5), 611–7. http://doi.org/10.1111/j.1467-7687.2012.01160.x

Bennett, J. (1978). Some remarks about concepts. *Behavioral and Brain Sciences*, *1*(4), 557–560.

Bialystok, E., Craik, F. I. M., & Luk, G. (2012). Bilingualism: consequences for mind and brain. *Trends in Cognitive Sciences*, *16*(4), 240–50. http://doi.org/10.1016/j.tics.2012.03.001

Biro, S., & Leslie, A. M. (2007). Infants' perception of goal-directed actions: development through cue-based bootstrapping. *Developmental Science*, *10*(3), 379–98. http://doi.org/10.1111/j.1467-7687.2006.00544.x

Bloom, P., & German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, *77*(1), 25–31.

Bugnyar, T., & Heinrich, B. (2005). Ravens, Corvus corax, differentiate between knowledgeable and ignorant competitors. *Proceedings. Biological Sciences / The Royal Society*, *272*(1573), 1641–6. http://doi.org/10.1098/rspb.2005.3144

Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition*, *112*(2), 337–42. http://doi.org/10.1016/j.cognition.2009.05.006

Buttelmann, F., Suhrke, J., & Buttelmann, D. (2015). What you get is what you believe: eighteen-month-olds demonstrate belief understanding in an unexpected-identity task. *Journal of Experimental Child Psychology*, *131*, 94–103. http://doi.org/10.1016/j.jecp.2014.11.009

Butterfill, S. A., & Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind and Language*, *28*(5), 606–637. http://doi.org/10.1111/mila.12036

Cacchione, T., Schaub, S., & Rakoczy, H. (2013). Fourteen-month-old infants infer the continuous identity of objects on the basis of nonvisible causal properties. *Developmental Psychology*, *49*(7), 1325–9. http://doi.org/10.1037/a0029746

Calero, C. I., Salles, A., Semelman, M., & Sigman, M. (2013). Age and gender dependent development of Theory of Mind in 6- to 8-years old children. *Frontiers in Human Neuroscience*, *7*(June), 281. http://doi.org/10.3389/fnhum.2013.00281

Call, J. (2001). Chimpanzee social cognition. *Trends in Cognitive Sciences*, *5*(9), 388–393.

Call, J., & Tomasello, M. (1999). A nonverbal false belief task: the performance of children and great apes. *Child Development*, *70*(2), 381–95. http://doi.org/10.1111/1467-8624.00028

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, *12*(5), 187–92. http://doi.org/10.1016/j.tics.2008.02.010

Callaghan, T., Rochat, P., Lillard, A., Claux, M. L., Odden, H., Itakura, S., … Singh, S. (2005). Synchrony in the onset of mental-state reasoning: Evidence from five cultures. *Psychological Science*, *16*(5), 378–384. http://doi.org/10.1111/j.0956-7976.2005.01544.x

Cannon, E. N., & Woodward, A. L. (2012). Infants generate goal-based action predictions. *Developmental Science*, *15*(2), 292–8. http://doi.org/10.1111/j.1467-7687.2011.01127.x

Carey, S. (2011). Précis of "The Origin of Concepts". *The Behavioral and Brain Sciences*, *34*(3), 113-24-62. http://doi.org/10.1017/S0140525X10000919

Carey, S., & Johnson, S. C. (2000). Knowledge enrichment and conceptual change in folkbiology: evidence from Williams syndrome. In *Sperber (ed.) Metarepresentations: A Multidisciplinary Perspective* (pp. 225–264).

Carey, S., & Xu, F. (2001). Infants' knowledge of objects: beyond object files and object tracking. *Cognition*, *80*(1–2), 179–213.

Carlson, S. M., Moses, L. J., & Claxton, L. J. (2004). Individual differences in executive functioning and theory of mind: An investigation of inhibitory control and planning ability. *Journal of Experimental Child Psychology*, *87*(4), 299–319. http://doi.org/10.1016/j.jecp.2004.01.002

Caron, A. J. (2009). Comprehension of the representational mind in infancy. *Developmental Review*, *29*(2), 69–95. http://doi.org/10.1016/j.dr.2009.04.002

Carruthers, P. (2013). Mindreading in Infancy. *Mind & Language*, *28*(2), 141–172. http://doi.org/10.1111/mila.12014

Carruthers, P. (2015a). Mindreading in adults: evaluating two-systems views. *Synthese*. http://doi.org/10.1007/s11229-015-0792-3

Carruthers, P. (2015b). Two Systems for Mindreading? *Review of Philosophy and*

Psychology, *7*(1), 141–162. http://doi.org/10.1007/s13164-015-0259-y

Cheng, M., Wang, L., & Leslie, A. M. (2016). Tracking multiple minds : Working memory (WM) and Executive Function (EF) in preschool theory of mind ., (January), 922184.

Chow, V., Poulin-Dubois, D., & Lewis, J. (2008). To see or not to see: Infants prefer to follow the gaze of a reliable looker. *Developmental Science*, *11*(5), 761–770. http://doi.org/10.1111/j.1467-7687.2008.00726.x

Christensen, W., & Michael, J. (2016). From two systems to a multi-systems architecture for mindreading. *New Ideas in Psychology*, *40*, 48–64. http://doi.org/10.1016/j.newideapsych.2015.01.003

Clements, W. A., & Perner, J. (1994). Implicit understanding of Belief. *Cognitive Development*, *395*, 377–395.

Clements, W. A., Rustin, C. L., & McCallum, S. (2000). Promoting the transirion from implicit to explicit understanding: a training study of false belief. *Developmental Science*, *3*(1), 81–92.

Cohen, A. S., & German, T. C. (2009). Encoding of others' beliefs without overt instruction. *Cognition*, *111*(3), 356–363. http://doi.org/10.1016/j.cognition.2009.03.004

Cordes, S., & Brannon, E. M. (2008). Quantitative competencies in infancy. *Developmental Science*, *11*(6), 803–8. http://doi.org/10.1111/j.1467-7687.2008.00770.x

Cowan, N. (2001). The magical number 4 in short term memory. A reconsideration of storage capacity. *Behavioral and Brain Sciences*, *24*(4), 87–186.

Csibra, G. (2007). Action mirroring and action understanding : an alternative account. In P. Haggard, Y. Rosetti, & M. Kawato (Eds.), *Attention and Perfomnace XXII: Sensorimotor Foundations of Higher Cognition* (pp. 435–459). Oxford University Press. http://doi.org/10.1093/acprof:oso/9780199231447.003.0020

Csibra, G. (2008). Goal attribution to inanimate agents by 6.5-month-old infants. *Cognition*, *107*(2), 705–17. http://doi.org/10.1016/j.cognition.2007.08.001

Csibra, G., Gergely, G., Biro, S., Koos, O., & Brockbank, M. (1999). Goal attribution without agency cues: The perception of "pure reason" in infancy. *Cognition*, *72*(3), 237–267. http://doi.org/10.1016/S0010-0277(99)00039-6

Csibra, G., Hernik, M., Mascaro, O., Tatone, D., & Lengyel, M. (2016). Statistical treatment of looking-time data. *Developmental Psychology*, *52*(4), 521–536. http://doi.org/10.1037/dev0000083

Csibra, G., Tucker, L. A., Volein, A., & Johnson, M. H. (2000). Cortical development and saccade planning: the ontogeny of the spike potential. *Neuroreport*, *11*(5), 1069–73. http://doi.org/10.1097/00001756-200004070-00033

Daum, M. M., Attig, M., Gunawan, R., Prinz, W., & Gredebäck, G. (2012). Actions seen through babies' eyes: A dissociation between looking time and predictive gaze. *Frontiers in Psychology*, *3*(SEP), 1–13. http://doi.org/10.3389/fpsyg.2012.00370

De Bruin, L. C., & Newen, A. (2012). An association account of false belief understanding. *Cognition*, *123*(2), 240–59. http://doi.org/10.1016/j.cognition.2011.12.016

Dennett, D. C. (1978). Beliefs about beliefs [P&W, SR&B]. *Behavioral and Brain Sciences*, *1*(4), 568–570.

EEGLAB Tutorial. (2001). Retrieved September 3, 2016, from https://sccn.ucsd.edu/wiki/EEGLAB#EEGLAB_Tutorial

Elekes, F., Varga, M., & Király, I. (2016). Evidence for spontaneous level-2 perspective taking in adults. *Consciousness and Cognition*, *41*, 93–103. http://doi.org/10.1016/j.concog.2016.02.010

Evans, J. S. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7*(10), 454–459. http://doi.org/10.1016/j.tics.2003.08.012

Fabricius, W. V., Boyer, T. W., Weimer, A. A., & Carroll, K. (2010). True or false: Do 5-year-olds understand belief? *Developmental Psychology*, *46*(6), 1402–1416. http://doi.org/10.1037/a0017648

Feigenson, L., & Carey, S. (2003). Tracking individuals via object-files: Evidence from infants' manual search. *Developmental Science*, *6*(5), 568–584. http://doi.org/10.1111/1467-7687.00313

Feigenson, L., & Carey, S. (2005). On the limits of infants' quantification of small object arrays. *Cognition*, *97*(3), 295–313. http://doi.org/10.1016/j.cognition.2004.09.010

Feigenson, L., Carey, S., & Hauser, M. (2002). The representations underlying infants' choice of more: Object files versus analog magnitudes. *Psychological Science*, *13*(2), 150–156. http://doi.org/10.1111/1467-9280.00427

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307–314. http://doi.org/10.1016/j.tics.2004.05.002

Feigenson, L., & Halberda, J. (2004). Infants chunk object arrays into sets of individuals. *Cognition*, *91*(2), 173–190. http://doi.org/10.1016/j.cognition.2003.09.003

Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children's knowledge about visual perception: Further evidence for the Level 1-Level 2 distinction. *Developmental Psychology*, *17*(1), 99–103. http://doi.org/10.1037/0012-1649.17.1.99

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *358*(1431), 459–473. http://doi.org/10.1098/rstb.2002.1218

Futó, J., Téglás, E., Csibra, G., & Gergely, G. (2010). Communicative function demonstration induces kind-based artifact representation in preverbal infants. *Cognition*, *117*(1), 1–8. http://doi.org/10.1016/j.cognition.2010.06.003

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mindreading. *Trends in Cognitive Sciencess in Cognitive Sciences*, *2*(12), 493–501. http://doi.org/10.1016/S1364-6613(98)01262-5

Garnham, W. A., & Ruffman, T. (2001). Doesn't see, doesn't know: is anticipatory looking really related to understanding or belief? *Developmental Science*, *4*(1), 94–100. http://doi.org/10.1111/1467-7687.00153

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: the naïve theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287–292. http://doi.org/10.1016/S1364-6613(03)00128-1

Gergely, G., Nádasdy, Á., Csibra, G., & Biro, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*(2), 165–193. http://doi.org/10.1016/0010-0277(95)00661-H

Gopnik, A., & Astington, J. W. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child Development*, *59*(1), 26–37. http://doi.org/10.2307/1130386

Gopnik, A., & Wellman, H. M. (1992). Why the Child's Theory of Mind Really Is a Theory. *Mind {&} Language*, *7*(1–2), 145–171. http://doi.org/10.1111/j.1468-0017.1992.tb00202.x

Goupil, L., Romand-monnier, M., & Kouider, S. (2016). Infants ask for help when they know they don't know. *Proceedings of the National Academy of Sciences*, *113*(13), 3492–3496. http://doi.org/10.1073/pnas.1515129113

Griffiths, D., Dickinson, A., & Clayton, N. (1999). Episodic memory: What can animals remember about their past? *Trends in Cognitive Sciences*, *3*(2), 74–80. http://doi.org/10.1016/S1364-6613(98)01272-8

Gweon, H., Dodell-Feder, D., Bedny, M., & Saxe, R. (2012). Theory of mind performance in children correlates with functional specialization of a brain region for thinking about thoughts. *Child Development*, *83*(6), 1853–68. http://doi.org/10.1111/j.1467-8624.2012.01829.x

Harman, G. (1978). Studying the chimpanzee's theory of mind. *Behavioral and Brain Sciences*, *1*(4), 576–577.

Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General*, *108*(3), 356–388.

Hassler, U., Barreto, N. T., & Gruber, T. (2011). Induced gamma band responses in human EEG after the control of miniature saccadic artifacts. *NeuroImage*, *57*(4), 1411–21. http://doi.org/10.1016/j.neuroimage.2011.05.062

Hauser, M. D., Carey, S., & Hauser, L. B. (2000). Spontaneous number representation in semi-free-ranging rhesus monkeys. *Proceedings of the Royal Society B: Biological Sciences*, *267*(1445), 829–33. http://doi.org/10.1098/rspb.2000.1078

Helming, K. A., Strickland, B., & Jacob, P. (2014). Making sense of early false-belief understanding. *Trends in Cognitive Sciences*, *18*(4), 167–170. http://doi.org/10.1016/j.tics.2014.01.005

Hernik, M., & Southgate, V. (2012). Nine-months-old infants do not need to know what the agent prefers in order to reason about its goals: On the role of preference and persistence in infants' goal-attribution. *Developmental Science*, *15*(5), 714–722. http://doi.org/10.1111/j.1467-7687.2012.01151.x

Herrmann, E., Call, J., Hernàndez-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: the cultural intelligence hypothesis. *Science*, *317*(5843), 1360–6. http://doi.org/10.1126/science.1146282

Heyes, C. M. (2014a). False belief in infancy: a fresh look. *Developmental Science*, *17*(5), 647–59. http://doi.org/10.1111/desc.12148

Heyes, C. M. (2014b). Submentalizing: I Am Not Really Reading Your Mind. *Perspectives on Psychological Science*, *9*(2), 131–143. http://doi.org/10.1177/1745691613518076

Heyes, C. M., & Frith, C. D. (2014). The cultural evolution of mind reading. *Science*, *344*(6190), 1243091. http://doi.org/10.1126/science.1243091

Hogrefe, G., Wimmer, H., & Perner, J. (1986). Ignorance versus False Belief : A Developmental Lag in Attribution of Epistemic States. *Child Development*, *57*(3), 567–582.

Hyde, D. C., Aparicio Betancourt, M., & Simon, C. E. (2015). Human temporal-parietal junction spontaneously tracks others' beliefs: A functional near-infrared spectroscopy study. *Human Brain Mapping*, *36*(12), 4831–46. http://doi.org/10.1002/hbm.22953

Izard, V., Sann, C., Spelke, E. S., & Streri, A. (2009). Newborn infants perceive abstract numbers. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(25), 10382–5. http://doi.org/10.1073/pnas.0812142106

Jacob, P. (2005). First-person and third-person mindreading. In *P. Gampieri-Deutsch (ed.) Psychoanalysis as an Empirical, Interdisciplinary Science. Austrian Academy of Sciences (2005)* (pp. 1–29).

Jacob, P. (2013). A puzzle about belief-ascription. In *Mind and Society: Cognitive Science meets the Social Sciences* (pp. 1–32).

Jacob, P. (2014). Why Reading Minds Is Not Like Reading Words.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*. http://doi.org/10.1037/0033-2909.114.1.3

Kahneman, D., Treisman, a, & Gibbs, B. J. (1992). The reviewing of object files: object-specific integration of information. *Cognitive Psychology*, *24*(2), 175–219.

Kampis, D., Fogd, D., & Kovács, Á. M. (n.d.). Nonverbal components of Theory of Mind in typical and atypical development. *Infant Behavior and Development*.

Kampis, D., Parise, E., Csibra, G., & Kovács, Á. M. (2015). Neural signatures for sustaining object representations attributed to others in preverbal human infants. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1819), 20151683-. http://doi.org/10.1098/rspb.2015.1683

Kampis, D., Somogyi, E., Itakura, S., & Király, I. (2013). Do infants bind mental states to agents? *Cognition*, *129*(2), 232–40. http://doi.org/10.1016/j.cognition.2013.07.004

Kaufman, J., Csibra, G., & Johnson, M. H. (2003). Representing occluded objects in the human infant brain. *Proceedings of the Royal Society B: Biological Sciences*, *270 Suppl*, S140-3. http://doi.org/10.1098/rsbl.2003.0067

Kaufman, J., Csibra, G., & Johnson, M. H. (2005). Oscillatory activity in the infant brain reflects object maintenance. *PNAS*, *102*(42), 15271–15274.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, *89*(1), 25–41. http://doi.org/10.1016/S0010-0277(03)00064-7

Kilner, J. M., Vargas, C., Duval, S., Blakemore, S.-J., & Sirigu, A. (2004). Motor activation prior to observation of a predicted movement. *Nature Neuroscience*, *7*(12), 1299–301. http://doi.org/10.1038/nn1355

Knudsen, B., & Liszkowski, U. (2011). 18-Month-Olds Predict Specific Action Mistakes Through Attribution of False Belief, Not Ignorance, and Intervene Accordingly. *Infancy*, no-no. http://doi.org/10.1111/j.1532-7078.2011.00105.x

Knudsen, B., & Liszkowski, U. (2012a). 18-Month-Olds Predict Specific Action Mistakes Through Attribution of False Belief, Not Ignorance, and Intervene Accordingly. *Infancy*, *17*(6), 672–691. http://doi.org/10.1111/j.1532-7078.2011.00105.x

Knudsen, B., & Liszkowski, U. (2012b). Eighteen- and 24-month-old infants correct others in anticipation of action mistakes. *Developmental Science*, *15*(1), 113–22. http://doi.org/10.1111/j.1467-7687.2011.01098.x

Koos, O., Gergely, G., Csibra, G., & Biro, S. (1997). Why eating Smarties makes you smart: Understanding of false belief at the age of 3. Poster presented at the biennial meeting of the society for research in child development, Washington, DC (April). In *Poster presented at the biennial meeting of the society for research in child development, Washington, DC (April)*.

Kornell, N. (2009). Metacognition in humans and animals. *Current Directions in Psychological Science*, *18*(1), 11–15. http://doi.org/10.1111/j.1467-8721.2009.01597.x

Köster, M. (2016). What about microsaccades in the electroencephalogram of infants? *Proceedings Oc the Royal Sciety: Biological Sciences*, *283*, 1835.

Köster, M., Friese, U., Schöne, B., Trujillo-Barreto, N., & Gruber, T. (2014). Theta-gamma coupling during episodic retrieval in the human EEG. *Brain Research*, *1577*, 57–68. http://doi.org/10.1016/j.brainres.2014.06.028

Kovács, Á. M. (2009). Early bilingualism enhances mechanisms of false-belief reasoning. *Developmental Science*, *12*(1), 48–54. http://doi.org/10.1111/j.1467-7687.2008.00742.x

Kovács, Á. M. (2015). Belief Files in Theory of Mind Reasoning. *Review of Philosophy and Psychology*, *7*(2), 509–527. http://doi.org/10.1007/s13164-015-0236-5

Kovács, Á. M., Kühn, S., Gergely, G., Csibra, G., & Brass, M. (2014). Are all beliefs equal? Implicit belief attributions recruiting core brain regions of theory of mind. *PloS One*, *9*(9), e106558. http://doi.org/10.1371/journal.pone.0106558

Kovács, Á. M., Tauzin, T., Téglás, E., Gergely, G., & Csibra, G. (2014). Pointing as epistemic request: 12-month-olds point to receive new information. *Infancy*, *19*(6), 543–557. http://doi.org/10.1111/infa.12060

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: susceptibility to others' beliefs in human infants and adults. *Science (New York, N.Y.)*, *330*(6012), 1830–4. http://doi.org/10.1126/science.1190792

Kuhlmeier, V., Wynn, K., Bloom, P., Kuhlmeier, V., Wynn, K., & Bloom, P. (2015). Attribution of Dispositional States By 12-Month-Olds, *14*(5), 402–408.

Leekam, S. (2016). Social cognitive impairment and autism: what are we trying to explain? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *371*(1686), 20150082. http://doi.org/10.1098/rstb.2015.0082

Leslie, A. M. (1987). Pretense and representation: The origins of "theory of mind." *Psychological Review*, *94*(4), 412–426. http://doi.org/10.1037//0033-295X.94.4.412

Leslie, A. M. (1994). Pretending and believing: issues in the theory of ToMM. *Cognition*, *50*(1–3), 211–38.

Leslie, A. M. (2000a). How to acquire a "representational theory of mind." In *Metarepresentation* (pp. 197–223).

Leslie, A. M. (2000b). "Theory of Mind" as a Mechanism of Selective Attention. In M. S. Gazzaniga (Ed.), *The new cognitive neurosciences* (2nd Editio, pp. 1235–1247). MIT Press. http://doi.org/10.1162/jocn.1995.7.4.514

Leslie, A. M., German, T. P., & Polizzi, P. (2005). Belief-desire reasoning as a process of selection. *Cognitive Psychology*, *50*(1), 45–85. http://doi.org/10.1016/j.cogpsych.2004.06.002

Leslie, A. M., & Thaiss, L. (1992). Domain specificity in conceptual development -

Neuropsychological evidence from autism, *43*, 225–251.

Leslie, A. M., Xu, F., Tremoulet, P. D., & Scholl, B. J. (1998). Indexing and the object concept: developing `what' and `where' systems. *Trends in Cognitive Sciences*, *2*(1), 10–8.

Leung, S., Mareschal, D., Rowsell, R., Simpson, D., Iaria, L., Grbic, A., & Kaufman, J. (2016). Oscillatory Activity in the Infant Brain and the Representation of Small Numbers. *Frontiers in Systems Neuroscience*, *10*(February), 1–7. http://doi.org/10.3389/fnsys.2016.00004

Liszkowski, U. (2013). Using Theory of Mind. *Child Development Perspectives*, *7*(2), 104–109. http://doi.org/10.1111/cdep.12025

Liszkowski, U., Carpenter, M., Striano, T., & Tomasello, M. (2006). 12- and 18-Month-Olds Point to Provide Information for Others. *Journal of Cognition and Development*, *7*(2), 173–187. http://doi.org/10.1207/s15327647jcd0702_2

Liu, D., Sabbagh, M. a, Gehring, W. J., & Wellman, H. M. (2004). Decoupling beliefs from reality in the brain: an ERP study of theory of mind. *Neuroreport*, *15*(6), 991–995. http://doi.org/10.1097/00001756-200404290-00012

Liu, D., Wellman, H. M., Tardif, T., & Sabbagh, M. a. (2008). Theory of mind development in Chinese children: a meta-analysis of false-belief understanding across cultures and languages. *Developmental Psychology*, *44*(2), 523–31. http://doi.org/10.1037/0012-1649.44.2.523

Low, J., Apperly, I. A., Butterfill, S. A., & Rakoczy, H. (2016). Cognitive Architecture of Belief Reasoning in Children and Adults: A Primer on the Two-Systems Account. *Child Development Perspectives*, *0*(0), 1–6. http://doi.org/10.1111/cdep.12183

Low, J., Drummond, W., Walmsley, A., & Wang, B. (2014). Representing how rabbits quack and competitors act: limits on preschoolers' efficient ability to track perspective. *Child Development*, *85*(4), 1519–34. http://doi.org/10.1111/cdev.12224

Low, J., & Watts, J. (2013). Attributing false beliefs about object identity reveals a signature blind spot in humans' efficient mind-reading system. *Psychological Science*, *24*(3), 305–11. http://doi.org/10.1177/0956797612451469

Luo, Y. (2011). Do 10-month-old infants understand others' false beliefs? *Cognition*, 1–29.

Luo, Y., & Baillargeon, R. (2005). Can a self-propelled box have a goal? Psychological reasoning in 5-month-old infants. *Psychological Science*, *16*(8), 601–608.

Luo, Y., & Baillargeon, R. (2007). Do 12.5-month-old infants consider what objects others can see when interpreting their actions? *Cognition*, *105*(3), 489–512. http://doi.org/10.1016/j.cognition.2006.10.007

Martin, A., & Santos, L. R. (2016). What Cognitive Representations Support Primate Theory of Mind? *Trends in Cognitive Sciences*, *20*(5), 375–382. http://doi.org/10.1016/j.tics.2016.03.005

McCleery, J. P., Surtees, A. D. R., Graham, K. A., Richards, J. E., & Apperly, I. A. (2011). The Neural and Cognitive Time Course of Theory of Mind. *Journal of Neuroscience*, *31*(36), 12849–12854. http://doi.org/10.1523/JNEUROSCI.1392-11.2011

McCrink, K., & Wynn, K. (2004). Large-number addition and subtraction by 9-month-old infants. *Psychological Science*, *15*(11), 776–81. http://doi.org/10.1111/j.0956-7976.2004.00755.x

McCrink, K., & Wynn, K. (2007). Ratio abstraction by 6-month-old infants. *Psychological Science*, *18*(8), 740–745. http://doi.org/10.1111/j.1467-9280.2007.01969.x

Melloni, L., Schwiedrzik, C. M., Wibral, M., Rodriguez, E., & Singer, W. (2009). Response to: Yuval-Greenberg et al., "Transient Induced Gamma-Band Response in EEG as a Manifestation of Miniature Saccades." Neuron 58, 429-441. *Neuron*, *62*(1), 8-10-12. http://doi.org/10.1016/j.neuron.2009.04.002

Merritt, D. J., & Brannon, E. M. (2011). Nothing to it: Precursors to a Zero Concept in Preschoolers. *Behavioral Processes*, *4*(164), 91–97. http://doi.org/10.1126/scisignal.2001449.Engineering

Miller, G. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, *101*(2), 343–352. http://doi.org/10.1037/h0043158

Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and theory of mind: meta-analysis of the relation between language ability and false-belief understanding. *Child Development*, *78*(2), 622–46. http://doi.org/10.1111/j.1467-8624.2007.01018.x

Mody, S., & Carey, S. (2016). The emergence of reasoning by the disjunctive syllogism in early childhood. *Cognition*, *154*, 40–48. http://doi.org/10.1016/j.cognition.2016.05.012

Moll, H., Meltzoff, A. N., Merzsch, K., & Tomasello, M. (2013). Taking versus confronting visual perspectives in preschool children. *Developmental Psychology*, *49*(4), 646–54. http://doi.org/10.1037/a0028633

Newcombe, N. S., Lloyd, M. E., & Ratliff, K. R. (2007). *Development Of Episodic And Autobiographical Memory: A Cognitive Neuroscience Perspective*. *Advances in Child Development and Behavior* (Vol. 35). http://doi.org/10.1016/B978-0-12-009735-7.50007-4

Nordmeyer, A. E., & Frank, M. C. (2014). The role of context in young children's comprehension of negation. *Journal of Memory and Language*, *77*(C), 25–39. http://doi.org/10.1016/j.jml.2014.08.002

Oktay-Gür, N., & Rakoczy, H. (2016). Why do children (even 6 - to 8 - year - olds) fail true belief tasks ? In *BCCCD Budapest*.

Okuyama, S., Kuki, T., & Mushiake, H. (2015). Representation of the Numerosity "zero" in the Parietal Cortex of the Monkey. *Scientific Reports*, *5*, 10059. http://doi.org/10.1038/srep10059

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*(5719), 255–8. http://doi.org/10.1126/science.1107621

Penn, D. C., & Povinelli, D. J. (2007). On the lack of evidence that non-human animals possess anything remotely resembling a "theory of mind". *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *362*(1480), 731–44. http://doi.org/10.1098/rstb.2006.2023

Perner, J. (2016). Mental Files in Development: A cognitive theory of how children represent belief. In *Understanding Communication and Understanding Minds: The Role of Metarepresentation. 27. June - 2. July, Central European University, Budapest*.

Perner, J., Huemer, M., & Leahy, B. (2015). Mental files and belief : A cognitive theory of how children represent belief and its intensionality, *145*, 77–88.

Perner, J., & Leahy, B. (2016). Mental Files in Development: Dual Naming, False Belief, Identity and Intensionality. *Review of Philosophy and Psychology*, *7*, 491–508. http://doi.org/10.1007/s13164-015-0235-6

Perner, J., Leekam, S. R., & Wimmer, H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology*, *5*, 125–137. http://doi.org/10.1111/j.2044-835X.1987.tb01048.x

Perner, J., Mauer, M. C., & Hildenbrand, M. (2011). Identity: key to children's understanding of belief. *Science*, *333*(6041), 474–7. http://doi.org/10.1126/science.1201216

Perner, J., & Ruffman, T. (2005). Infants ' Insight into the Mind: how deep? *Science*, (April), 214–216.

Perner, J., Ruffman, T., & Leekam, S. R. (1994). Theory of Mind Is Contagious: You Catch It from Your Sibs. *Child Development*, *65*(4), 1228–1238. http://doi.org/10.1111/j.1467-8624.1994.tb00814.x

Perner, J., Stummer, S., Sprung, M., & Doherty, M. (2002). Theory of mind finds its Piagetian perspective: Why alternative naming comes with understanding belief. *Cognitive Development*, *17*(3–4), 1451–1472. http://doi.org/10.1016/S0885-2014(02)00127-2

Poulin-Dubois, D., & Chow, V. (2009). The effect of a looker's past reliability on infants' reasoning about beliefs. *Developmental Psychology*, *45*(6), 1576–1582. http://doi.org/10.1037/a0016715

Povinelli, D. J., & Vonk, J. (2003). Chimpanzee minds: suspiciously human? *Trends in Cognitive Sciences*, *7*(4), 157–160. http://doi.org/10.1016/S1364-6613(03)00053-6

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

Pylyshyn, Z. W. (1978). When is attribution of beliefs justified?[P&W]. *Behavioral and Brain Sciences*, *1*(4), 592–593.

Pylyshyn, Z. W. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition*, *80*(1–2), 127–158. http://doi.org/10.1016/S0010-0277(00)00156-6

Quine, W. V. O. (1961). *From a logical point of view: 9 logico-philosophical essays*. Harvard University press.

Qureshi, A. W., Apperly, I. A., & Samson, D. (2010). Executive function is necessary for perspective selection, not Level-1 visual perspective calculation: Evidence from a dual-task study of adults. *Cognition*, *117*(2), 230–236. http://doi.org/10.1016/j.cognition.2010.08.003

Rakoczy, H. (2012). Do infants have a theory of mind? *The British Journal of Developmental Psychology*, *30*(Pt 1), 59–74. http://doi.org/10.1111/j.2044-835X.2011.02061.x

Rakoczy, H., Bergfeld, D., Schwarz, I., & Fizke, E. (2015). Explicit Theory of Mind Is Even More Unified Than Previously Assumed: Belief Ascription and Understanding Aspectuality Emerge Together in Development. *Child Development*, *86*(2), 486–502. http://doi.org/10.1111/cdev.12311

Recanati, F. (2012). *Mental Files*. *Mental Files*. Oxford University Press. http://doi.org/10.1093/acprof:oso/9780199659982.001.0001

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience*, *2*(9), 661–70. http://doi.org/10.1038/35090060

Rubio-Fernández, P. (2016). Belief reasoning in spontaneous conversation : The sky ' s the signature limit. In *Talk presented at Understanding Communication and Understanding Minds: The Role of Metarepresentation. 27. June - 2. July, Central European University, Budapest*.

Rubio-Fernández, P., & Geurts, B. (2013). How to pass the false-belief task before your 4th birthday. *Psychological Science*, *24*(1), 27–33. http://doi.org/10.1177/0956797612447819

Rueschemeyer, S.-A., Gardner, T., & Stoner, C. (2015). The Social N400 effect: how the presence of other listeners affects language comprehension. *Psychonomic Bulletin & Review*, *22*(1), 128–34. http://doi.org/10.3758/s13423-014-0654-x

Rueschemeyer, S., Gardner, T., & Stoner, C. (2014). The Social N400 effect: how the presence of other listeners affects language comprehension. *Psychonomic Bulletin & Review*, *22*(1), 128–134. http://doi.org/10.3758/s13423-014-0654-x

Ruffman, T., Slade, L., & Crowe, E. (2002). The relation between children's and mothers' mental state language and theory-of-mind understanding. *Child Development*, *73*(3), 734–751. http://doi.org/10.1111/1467-8624.00435

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Bodley Scott, S. E. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(5), 1255–1266. http://doi.org/10.1037/a0018729

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind." *NeuroImage*, *19*(4), 1835–1842. http://doi.org/10.1016/S1053-8119(03)00230-1

Saxe, R., Tzelnic, T., & Carey, S. (2007). Knowing who dunnit: Infants identify the causal agent in an unseen causal interaction. *Developmental Psychology*, *43*(1), 149–58. http://doi.org/10.1037/0012-1649.43.1.149

Saxe, R., & Wexler, A. (2005). Making sense of another mind: the role of the right temporo-parietal junction. *Neuropsychologia*, *43*(10), 1391–9. http://doi.org/10.1016/j.neuropsychologia.2005.02.013

Schlottmann, A., & Ray, E. (2010). Goal attribution to schematic animals: do 6-month-olds perceive biological motion as animate? *Developmental Science*, *13*(1), 1–10. http://doi.org/10.1111/j.1467-7687.2009.00854.x

Schneider, D., Bayliss, A. P., Becker, S. I., & Dux, P. E. (2011). Eye movements reveal sustained implicit processing of others' mental states. *Journal of Experimental Psychology. General*, *141*(3), 433–8. http://doi.org/10.1037/a0025458

Schneider, D., Bayliss, A. P., Becker, S. I., & Dux, P. E. (2012). Eye movements reveal sustained implicit processing of others' mental states. *Journal of Experimental Psychology. General*, *141*(3), 433–8. http://doi.org/10.1037/a0025458

Schneider, D., Lam, R., Bayliss, A. P., & Dux, P. E. (2012). Cognitive Load Disrupts Implicit Theory-of-Mind Processing. *Psychological Science*, *23*(8), 842–847. http://doi.org/10.1177/0956797612439070

Schneider, D., Slaughter, V. P., & Dux, P. E. (2015). What do we know about implicit false-belief tracking? *Psychonomic Bulletin & Review*, *22*(1), 1–12. http://doi.org/10.3758/s13423-014-0644-z

Scholl, B. J., & Leslie, A. M. (1999). Modularity , Development and " Theory of Mind ." *Mind & Language*, *14*(1), 131–153.

Scholl, B. J., & Pylyshyn, Z. W. (1999). Tracking multiple items through occlusion: clues to visual objecthood. *Cognitive Psychology*, *38*(2), 259–90. http://doi.org/10.1006/cogp.1998.0698

Scholl, B. J., Pylyshyn, Z. W., & Feldman, J. (2001). What is a visual object? Evidence from target merging in multiple object tracking. *Cognition*, *80*(1–2), 159–177. http://doi.org/10.1016/S0010-0277(00)00157-8

Scott, R. M., & Baillargeon, R. (2009). Which Penguin Is This ? Attributing False Beliefs About Object 18 Months. *Child Development*, *80*(4), 1172–1196.

Scott, R. M., & Baillargeon, R. (2014). How fresh a look? A reply to Heyes. *Developmental Science*, *17*(5), 660–664. http://doi.org/10.1111/desc.12173

Scott, R. M., He, Z., Baillargeon, R., & Cummins, D. (2012). False-belief understanding in 2.5-year-olds: evidence from two novel verbal spontaneous-response tasks. *Developmental Science*, *15*(2), 181–93. http://doi.org/10.1111/j.1467-7687.2011.01103.x

Scott, R. M., & Roby, E. (2015). Processing demands impact 3-year-olds' performance in a spontaneous-response task: New evidence for the processing-load account of early false-belief understanding. *PLoS ONE*, *10*(11), 1–20. http://doi.org/10.1371/journal.pone.0142405

Senju, A., Southgate, V., Snape, C., Leonard, M., & Csibra, G. (2011). Do 18-month-olds really attribute mental states to others? A critical test. *Psychological Science*, *22*(7), 878–80. http://doi.org/10.1177/0956797611411584

Setoh, P., Scott, R. M., & Baillargeon, R. (2011). False-belief reasoning in 2.5-year-olds: Evidence from an elicited-response low-inhibition task. In *Biennial Meeting of the Society for Research in Child Development, Montreal, Canada*.

Shahaeian, A., Peterson, C. C., Slaughter, V., & Wellman, H. M. (2011). Culture and the sequence of steps in theory of mind development. *Developmental Psychology*, *47*(5), 1239–1247. http://doi.org/10.1037/a0023899

Shiffrin, R. M., & Schneider, W. (1984). Automatic and controlled processing revisited. *Psychological Review*, *91*(2), 269–276. http://doi.org/10.1037/0033-295X.91.2.269

Shimizu, Y. A., & Johnson, S. C. (2004). Infants' attribution of a goal to a morphologically unfamiliar agent. *Developmental Science*, *7*(4), 425–430. http://doi.org/10.1111/j.1467-7687.2004.00362.x

Siegal, M., & Beattie, K. (1991). Where to look first for children's knowledge of fake liefs ". *Cognition*, *38*, 1–12.

Sodian, B., Licata, M., Kristen-Antonow, S., Paulus, M., Killen, M., & Woodward, A. L. (2016). Understanding of Goals, Beliefs, and Desires Predicts Morally Relevant Theory of Mind: A Longitudinal Investigation. *Child Development*, (April 2016). http://doi.org/10.1111/cdev.12533

Song, H., & Baillargeon, R. (2008). Infants' reasoning about others' false perceptions. *Developmental Psychology*, *44*(6), 1789–95. http://doi.org/10.1037/a0013774

Southgate, V., & Begus, K. (2013). Motor activation during the prediction of nonexecutable actions in infants. *Psychological Science*, *24*(6), 828–35. http://doi.org/10.1177/0956797612459766

Southgate, V., Chevallier, C., & Csibra, G. (2010). Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. *Developmental Science*, *13*(6), 1–6. http://doi.org/10.1111/j.1467-7687.2009.00946.x

Southgate, V., Johnson, M. H., & Csibra, G. (2008). Infants attribute goals even to biomechanically impossible actions. *Cognition*, *107*(3), 1059–69. http://doi.org/10.1016/j.cognition.2007.10.002

Southgate, V., Johnson, M. H., El Karoui, I., & Csibra, G. (2010). Motor system activation reveals infants' on-line prediction of others' goals. *Psychological Science*, *21*(3), 355–9. http://doi.org/10.1177/0956797610362058

Southgate, V., Johnson, M. H., Osborne, T., & Csibra, G. (2009). Predictive motor activation during action observation in human infants. *Biology Letters*, *5*(6), 769–72. http://doi.org/10.1098/rsbl.2009.0474

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, *18*(7), 587–92. http://doi.org/10.1111/j.1467-9280.2007.01944.x

Southgate, V., & Vernetti, A. (2014). Belief-based action prediction in preverbal infants. *Cognition*, *130*(1), 1–10. http://doi.org/10.1016/j.cognition.2013.08.008

Spelke, E. S. (1990). Principles of object perception. *Cognitive Science*, *14*(1), 29–56. http://doi.org/10.1016/0364-0213(90)90025-R

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

Spelke, E. S. (2000). Core Knowledge. *American Psychologist*, (November), 1233–1243. http://doi.org/10.1037/0003-066X.55.11.1233

Sperber, D. (2000). Metarepresentations in an evolutionary perspective.pdf. In D. Sperber (Ed.), *Metarepresentations: A multidisciplinary perspective* (pp. 117–137). Oxford University Press.

Sprung, M., Perner, J., & Mitchell, P. (2007). Opacity and discourse referents: Object identity and object properties. *Mind and Language*, *22*(3), 215–245. http://doi.org/10.1111/j.1468-0017.2007.00307.x

Stahl, A. E., & Feigenson, L. (2014). Social knowledge facilitates chunking in infancy. *Child Development*, *85*(4), 1477–1490. http://doi.org/10.1111/cdev.12217

Stroop, J. R. (1936). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662.

Suddendorf, T. (2012). The rise of the metamind. In *The Descent of Mind: Psychological Perspectives on Hominid Evolution*. http://doi.org/10.1093/acprof:oso/9780192632593.003.0012

Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science*, *18*(7), 580–6. http://doi.org/10.1111/j.1467-9280.2007.01943.x

Surian, L., & Geraci, A. (2012). Where will the triangle look for it? Attributing false beliefs to a geometric shape at 17 months. *The British Journal of Developmental Psychology*, *30*(Pt 1), 30–44. http://doi.org/10.1111/j.2044-835X.2011.02046.x

Surtees, A. D. R., & Apperly, I. A. (2012). Egocentrism and Automatic Perspective Taking in Children and Adults. *Child Development*, *83*(2), 452–460. http://doi.org/10.1111/j.1467-8624.2011.01730.x

Surtees, A. D. R., Apperly, I. A., & Samson, D. (2016). I've got your number: Spontaneous perspective-taking in an interactive task. *Cognition*, *150*, 43–52. http://doi.org/10.1016/j.cognition.2016.01.014

Surtees, A. D. R., Butterfill, S. A., & Apperly, I. A. (2012). Direct and indirect measures of Level-2 perspective-taking in children and adults. *British Journal of Developmental Psychology*, *30*(1), 75–86. http://doi.org/10.1111/j.2044-835X.2011.02063.x

Surtees, A. D. R., Samson, D., & Apperly, I. A. (2016). Unintentional perspective-taking calculates whether something is seen, but not how it is seen. *Cognition*, *148*, 97–105. http://doi.org/10.1016/j.cognition.2015.12.010

Thoermer, C., Sodian, B., Vuori, M., Perst, H., & Kristen, S. (2012). Continuity from an implicit to an explicit understanding of false belief from infancy to preschool age. *British Journal of Developmental Psychology*, *30*(1), 172–187. http://doi.org/10.1111/j.2044-835X.2011.02067.x

Tomasello, M., Call, J., & Hare, B. (2003). Chimpanzees understand psychological states – the question is which ones and to what extent. *Trends in Cognitive Sciences*, *7*(4), 153–156. http://doi.org/10.1016/S1364-6613(03)00035-4

Van de Walle, G. a., Carey, S., & Prevor, M. (2000). Bases for Object Individuation in Infancy: Evidence From Manual Search. *Journal of Cognition and Development*, *1*(3), 249–280. http://doi.org/10.1207/S15327647JCD0103_1

van der Wel, R. P. R. D., Sebanz, N., & Knoblich, G. (2014). Do people automatically track others' beliefs? Evidence from a continuous measure. *Cognition*, *130*(1), 128–33. http://doi.org/10.1016/j.cognition.2013.10.004

Van Overwalle, F., & Vandekerckhove, M. (2013). Implicit and explicit social mentalizing: dual processes driven by a shared neural network. *Frontiers in Human Neuroscience*, *7*(September), 560. http://doi.org/10.3389/fnhum.2013.00560

vanMarle, K. (2013). Infants use different mechanisms to make small and large number ordinal judgments. *Journal of Experimental Child Psychology*, *114*(1), 102–10. http://doi.org/10.1016/j.jecp.2012.04.007

Wang, L., & Leslie, A. M. (2013). False belief and working memory: can children represent two false beliefs spontaneously? In *talk presented at Budapest CEU Conference on Cognitive Development, Budapest*.

Wang, L., & Leslie, A. M. (2016). Is Implicit Theory of Mind the "Real Deal"? The Own-Belief/True-Belief Default in Adults and Young Preschoolers. *Mind & Language*, *31*(2), 147–176. http://doi.org/10.1111/mila.12099

Wellman, H. M. (1992). *The child's theory of mind. The MIT Press series in learning, development, and conceptual change.* http://doi.org/10.2307/3977400

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. *Child Development*, *72*(3), 655–84.

Wellman, H. M., Phillips, A. T., Dunphy-Lelii, S., & LaLonde, N. (2004). Infant social attention predicts preschool social cognition. *Developmental Science*, *7*(3), 283–288. http://doi.org/10.1111/j.1467-7687.2004.00347.x

Westra, E. (2016). Pragmatic development and the false belief task. *Review of Philosophy and Psychology*. http://doi.org/10.1007/s13164-016-0320-5

Wilcox, T. (1999). Object individuation: Infants' use of shape, size, pattern, and color. *Cognition*, *72*(2), 125–166. http://doi.org/10.1016/S0010-0277(99)00035-9

Wilson, D. (2000). Metarepresentation in linguistic communication. In D. Sperber (Ed.), *Metarepresentations: A multidisciplinary perspective* (Vol. 11, pp. 411–448). Oxford University Press. http://doi.org/10.1017/CBO9781139028370.014

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constrancing function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*, 103–128. http://doi.org/10.1016/0010-0277(83)90004-5

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1–34. http://doi.org/10.1016/S0010-0277(98)00058-4

Wynn, K. (1992). Addition and subtraction by human infants. *Nature*, *358*(6389), 749–750. http://doi.org/10.1038/358749a0

Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. *Cognition*, *85*(3), 223–250. http://doi.org/10.1016/S0010-0277(02)00109-9

Xu, F., & Arriaga, R. I. (2007). Number discrimination in 10-month-old infants. *British Journal of Developmental Psychology*, *25*(1), 103–108. http://doi.org/10.1348/026151005X90704

Xu, F., & Carey, S. (1996). Infants' metaphysics: The case of numerical identity. *Cognitive Psychology*, *30*(2), 111–153. http://doi.org/10.1006/cogp.1996.0005

Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, *74*(1), 1–11. http://doi.org/10.1016/S0010-0277(99)00066-9

Yott, J., & Poulin-Dubois, D. (2012). Breaking the rules: do infants have a true understanding of false belief? *The British Journal of Developmental Psychology*, *30*(Pt 1), 156–71. http://doi.org/10.1111/j.2044-835X.2011.02060.x

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(20), 8235–40. http://doi.org/10.1073/pnas.0701408104

Yuval-Greenberg, S., Tomer, O., Keren, A. S., Nelken, I., & Deouell, L. Y. (2008). Transient induced gamma-band response in EEG as a manifestation of miniature saccades. *Neuron*, *58*(3), 429–41. http://doi.org/10.1016/j.neuron.2008.03.027

Zaitchik, D. (1990). When representations conflict with reality: The preschooler's problem with false beliefs and "false" photographs. *Cognition*, *35*(1), 41–68. http://doi.org/10.1016/0010-0277(90)90036-J

Zmyj, N., Prinz, W., & Daum, M. M. (2015). Eighteen-month-olds' memory interference and distraction in a modified A-not-B task is not associated with their anticipatory looking in a false-belief task. *Frontiers in Psychology*, *6*(August 2016), 857. http://doi.org/10.3389/fpsyg.2015.00857

Zosh, J. M., & Feigenson, L. (2012). Memory load affects object individuation in 18-month-old infants. *Journal of Experimental Child Psychology*, *113*(3), 322–336. http://doi.org/10.1016/j.jecp.2012.07.005

ACKNOWLEDGEMENT

TO EXTERNAL FUNDING AGENCIES

CONTRIBUTING TO PHD DISSERTATIONS

Name of doctoral candidate: Dora Kampis

Title of dissertation: MINDREADERS IN THE CRIB: COGNITIVE MECHANISMS FOR REPRESENTING OTHERS' MENTAL STATES IN HUMAN INFANTS

Name of supervisor(s): Agnes M. Kovacs, Gergely Csibra

External funding agency: **European Research Council**

Acknowledgement:

This research has received funding from the European Research Council (ERC) under the European Union's Seventh Framework Programme(FP7/2007-2013) under **grant agreement No. 284236 - REPCOLLAB**