Humaniora

DISSERTATIONES
PHILOSOPHICAE
UNIVERSITATIS
TARTUENSIS
10

VIVIAN BOHL

How do we understand others? Beyond theories of mindreading and interactionism

DISSERTATIONES PHILOSOPHICAE UNIVERSITATIS TARTUENSIS
10

DISSERTATIONES PHILOSOPHICAE UNIVERSITATIS TARTUENSIS
10

VIVIAN BOHL

How do we understand others? Beyond theories of mindreading and interactionism

UNIVERSITY OF TARTU
PRESS
1632

Dissertation has been accepted for defence of the degree of Doctor of Philosophy (PhD) in Philosophy in May 26, 2014 by the Council of the Institute of Philosophy and Semiotics, University of Tartu

Supervisor: Dr. Bruno Mölder

Opponent: Prof. Dr. Marc V. P. Slors
(Radboud University Nijmegen, the Netherlands)

Defence: the thesis will be defended at the University of Tartu, Estonia, on August 22, 2014, at 13.00, in the Senate room

This thesis has been published with the support of European Union through the European Social Fund (Graduate School of Linguistics, Philosophy and Semiotics)

ISSN 1406-9520
ISBN 978-9949-32-606-8 (print)
ISBN 978-9949-32-607-5 (pdf)

University of Tartu Press
www.tyk.ee

We must therefore rediscover, after the natural world,
the social world, not as an object or sum of objects,
but as a permanent field or dimension of existence.

Maurice Merleau-Ponty

ACKNOWLEDGEMENTS

This thesis would not have become a reality without the encouragement of my family and friends and the inspiration and feedback that I have received from my teachers and colleagues. First of all, I would like to thank Bruno Mölder for throwing me into the ocean of contemporary interdisciplinary research and teaching me how to swim. His advice and support throughout the process of writing the thesis has been invaluable. I'm greatly indebted to my co-authors, Wouter van den Bos and Nivedita Gangopadhyay, who made the process of article writing an exciting and fulfilling experience. I thank Dan Zahavi for giving me the opportunity to spend a part of my doctoral studies at the Center for Subjectivity Research in Copenhagen, and Alan Page Fiske for inviting me to UCLA and for reading manuscripts of appendices 2 and 3. I'm very grateful to the members of the *European Platform* research group (Wouter van den Bos, Marion Godman, Mog Stapleton, Christoph Teufel, and Marijn van Wingerden) for all the interdisciplinary discussions and collaboration. I would also like to thank Külli Teder and Marion Godman, for proofreading appendix 1, and Alexander Stewart Davies, for proofreading the summary article. Finally, I would like to thank the Department of Philosophy at the University of Tartu for providing me with financial support and a stimulating research environment. My research has been financially supported by the Graduate School of Linguistics, Philosophy and Semiotics at the University of Tartu, the Archimedes Foundation, Volkswagen Stiftung, and the following grants: ETF9117, SF0180110s08, IUT20-5.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS 6

LIST OF ORIGINAL PUBLICATIONS 8

0. INTRODUCTORY REMARKS 9

1. THE ROLE OF PHILOSOPHY IN SOCIAL COGNITION RESEARCH .. 15
   - 1.1. Philosophy is not just "armchair speculation" 16
   - 1.2. Painting with broad brush strokes? 19
   - 1.3. Conceptual hygiene 21
   - 1.4. The role of phenomenology in social cognition research 23
   - 1.5. Conclusion 27

2. THE STANDARD APPROACH TO SOCIAL COGNITION: READING EACH-OTHER'S MINDS 29
   - 2.1. From Premack and Woodruff's chimpanzee experiments to the false belief tasks 29
   - 2.2. The mainstream theories of mindreading 34
   - 2.3. Theory-theory 36
     - 2.3.1. The Child-Scientist Theory 37
     - 2.3.2. The Modularist Theory 45
   - 2.4. Simulation theory 53
     - 2.4.1. The early versions of simulation theory: Heal and Gordon 54
     - 2.4.2. Varieties of simulation 56
     - 2.4.3. Alvin Goldman's account: From pure ST to a ST-TT hybrid 57
     - 2.4.4. Some general challenges for ST 64
   - 2.5. The hybrid account of Nichols and Stich 66
   - 2.6. Concluding remarks 69

3. AGAINST MINDREADING: THE INTERACTIONIST APPROACH 71
   - 3.1. Changing the explanandum 72
   - 3.2. The chicken-and-egg problem: Mindreading or social interaction? 75
   - 3.3. Embodied social cognition 78
   - 3.4. Second person versus third person mode of social cognition 80
   - 3.5. Social interaction as constitutive of social cognition 82
   - 3.6. Contextual understanding 85
   - 3.7. Concluding remarks 88

4. SUMMARY AND CONCLUSIONS 90

REFERENCES 95

PUBLICATIONS 105

SUMMARY IN ESTONIAN 171

CURRICULUM VITAE 177

ELULOOKIRJELDUS 181

LIST OF ORIGINAL PUBLICATIONS

I. Bohl, V., & van den Bos, W. (2012). Toward an integrative account of social cognition: Marrying theory of mind and interactionism to study the interplay of Type 1 and Type 2 processes. *Frontiers in Human Neuroscience*, *6*, 1–15. Published online: 11 October 2012.

II. Bohl, V., & Gangopadhyay, N. (2013). Theory of mind and the unobservability of other minds. *Philosophical Explorations*, 1–20. Published online: 30 July 2013.

III. Bohl, V. (2014). We read minds to shape relationships. *Philosophical Psychology*, 1–21. Published online: 12 March 2014.

0. INTRODUCTORY REMARKS

We, humans, are a hyper-social species. We are highly dependent upon each other. However, when we act together we are capable of creating things that no other species is capable of creating, such as culture, science, and technology. *Social cognition* – the ability to interact with and to understand others[1] – is crucial to almost every aspect of our lives. Most of the time, we are rather good at understanding the people around us and we can efficiently and often effortlessly coordinate our behaviour with our family, friends, colleagues, and even strangers. But how do we do this? What are the basic cognitive processes that enable humans to understand and to interact with each other? This is the main question of contemporary interdisciplinary research on social cognition. This question also drives my dissertation.

For more than 30 years, philosophers and psychologists have assumed that the key competence in human social cognition is *mindreading*: the ability to predict and explain other individuals' behaviour in terms of their *mental states* – perceptions, beliefs, desires, intentions, emotions, etc. The mainstream theories of mindreading are the *theory-theory*, the *simulation theory*, and some combinations of the two. Simply put, theory-theory says that when we mindread, we make inferences about other people's mental states by using an implicit theory-like body of knowledge consisting of law-like principles that describe how different types of mental state interconnect and causally link to behaviour. According to simulation theory, we read other people's minds by mentally putting ourselves in their shoes: we imagine what it would be like to be in the other person's position and then attribute the mental states that arise in us to that other person.

Recently, theories of mindreading – both theory-theory and simulation-theory – have come under serious attack by authors who have a background in phenomenology and/or in accounts of embodied, embedded, enactive, and extended cognition. The critics are not only criticising the specific explanations that the mainstream theories of mindreading provide, but they argue against the idea that mindreading is as central to human social cognition as the proponents of standard theories of mindreading have assumed. The critics emphasize that the central question should not be "How do people read minds?" but rather "What enables people to interact?" Many of the critics argue that most real life social interactions take place in the absence of mindreading and thus that mindreading plays no important role in human social cognition. I'm calling the latter approach *interactionism*.[2]

---

[1] I borrowed the definition of social cognition from Spaulding (2012, p. 431).

[2] John Michael (2011, pp. 559–560) has defined interactionism as referring to a family of positions that endorse the claim that "social understanding and interaction do not require mindreading because various embodied and/or extended capabilities sustain social understanding and interaction in the absence of mindreading."

My thesis focuses on the current debate between theories of mindreading and interactionism. I think that interactionists correctly point out that the underlying assumption of theories of mindreading, that mindreading is at the core of human social cognition and necessary for human social interaction, is unwarranted. The standard research on mindreading has been based on laboratory experiments that are specifically designed to elicit mindreading. However, such experiments cannot possibly tell us how commonly people mindread in real life or what motivates them to do so. Since the prevailing assumption of the standard mindreading paradigm has been that mindreading is ubiquitous in human social cognition[3], the question "When and why do people actually mindread?" has, unfortunately, not been posed. However, it would be premature to buy the opposite interactionist claim that mindreading is marginal and plays no essential role in human social cognition. The truth is that we simply do not know how much people actually rely on mindreading in real life social situations. It is time to do some serious research on this issue! It is time to take a critical, yet open-minded look at the current debate between theories of mindreading and interactionism.

My dissertation consists of a summary article, three theoretical research articles that have been published in international peer-reviewed journals (appendices 1–3), and another shorter summary in Estonian. The summary article begins with a chapter on methodological issues, where I discuss the specific role of philosophy in social cognition research. I analyse methods that philosophers can and should adopt when they are involved in interdisciplinary research on social cognition and I explain what methods I have been using in the research articles that form the main part of the thesis. The methodological chapter is followed by an overview of the current approaches to social cognition. Chapter 2 is dedicated to the mainstream theories of mindreading, each of which assumes that human social cognition is first and foremost a matter of attributing mental states to others: *theory-theory* and *simulation theory*. Chapter 3 introduces a more recent approach that challenges the mainstream view: *interactionism*.

As becomes apparent in the summary article, the current theoretical situation of the field has in recent years taken the form of a heated debate between theories of mindreading and interactionism, wherein one side assumes that mindreading is crucial for social cognition, and the other side argues for the opposite. I'm convinced that the way forward is to move beyond the black-or-white battle between theories of mindreading and interactionism. What we need is a broader perspective on social cognition that would include the study of mindreading as well as the study of non-mentalistic social cognitive abilities, and which would, among other things, enable us to study empirically when and why people actually engage in mindreading in everyday life.

The first research article (co-authored with the neuroscientist Wouter van den Bos, who is currently working at the Max Planck Institute for Human

[3] See Slors (2012) for a possible explanation of why mindreading *seems* ubiquitous.

Development in Berlin) is entitled "Towards an integrative account of social cognition: Marrying theory of mind and interactionism to study the interplay of Type 1 and Type 2 processes." It was published as part of the research topic "Towards a neuroscience of social interaction" in a special issue of *Frontiers of Human Neuroscience* in 2012. In the article, we use the term "theory of mind" as a label for the mainstream theories of mindreading – i.e. for both theory-theory and simulation theory. We argue that instead of treating the standard theories of mindreading and interactionism as mutually exclusive opponents, these two approaches should be integrated into a more comprehensive account of social cognition. We rely on dual process models of social cognition that differentiate between two types of social cognitive processing. Processes of the first type (labelled Type 1) are typically fast, stimulus-driven, relatively inflexible, and therefore also highly efficient. In contrast, processes of the second type (labelled Type 2) are relatively slow and cognitively effortful, but at the same time allow for more flexibility, and may involve some conscious control. We argue that while interactionism focuses on aspects of social cognition that are mostly related to Type 1 processes, theories of mindreading typically focus on aspects of social cognition driven by Type 2 processes. We argue that it is plausible that in the majority of real life social interactions, both types of process are involved and that social cognition and behaviour may be sustained by the interplay between them. Finally, we look at how the new integrative theoretical framework can support experimental research on social interaction.

As the first author of the article, I am responsible for working out the overall idea of the paper and for doing most of the writing. My co-author Wouter van den Bos took care of the empirical details, especially with respect to the parts of the paper wherein we review the relevant neuroscientific studies and neuroscientific methods. His support was a great help throughout the whole process of writing and publishing the paper. Collaborating on the paper was a truly inspiring interdisciplinary experience. The most valuable contribution of the article derives from its theoretical content. Theoretical considerations led us to a new empirically testable hypothesis that Type 1 and Type 2 processes may mutually influence each-other during social interactions. I and my co-author are members of an interdisciplinary research group which forms part of the *European Platform for Life Sciences, Mind Sciences, and the Humanities*. Our group is currently designing pilot-experiments to test the above described hypothesis.

The article received a commentary by Hanne De Jaegher and Ezequiel Di Paolo (2013), who consider "Bohl and van den Bos's proposal as a research heuristic that can surely enrich empirical data" (p. 2) and estimate the integrative account as having "empirical potential" (p. 1). De Jaegher and Di Paolo outline two possible interpretations of our integrative proposal: "1. Individual sense-making is largely supported by Type 2 processes and interaction by Type1 processes" and "2. The relation between individual sense-making and interactive performances is analogous to the relation between Type 2 and Type 1

processes." They claim that since the first interpretation is implausible (it is indeed explicitly denied by myself and van den Bos as we argue that social interaction involves an interplay of both types of process), the second interpretation must be correct. However, this is not what we had in mind either. Our aim was to argue that interactionist accounts have been focusing largely on Type 1 processes, whereas theories of mindreading have been mainly targeting Type 2 processes, and to hypothesize that both types of process are involved in real life social interactions. This in turn was meant to highlight the necessity of a broader theoretical framework that would enable the study of both types of process and would include insights from both, theories of mindreading and interactionism. Our approach does not, however, translate into claim 2.

De Jaegher and Di Paolo also point out that their own position is not as radical as portrayed in our article, where it is classified as "interactionist." They explain that they do not assume that interactive factors can account for all social cognition or that supra-individual explanations can replace sub-personal explanations. It is good to have this issue clarified because it shows that our views converge in this respect. However, in the research agenda of De Jaegher and Di Paolo, no essential role is given to mindreading. This is where our views really do come apart. Unlike De Jaegher and Di Paolo, I think that mindreading, although not necessary for all social cognition, plays a specific role in human social understanding and that the standard theories of mindreading have the potential to partly explain how human social cognition works.

The second research article (co-authored with the philosopher Nivedita Gangopadhyay who currently works at the Ruhr-University Bochum) carries the title "Theory of mind and the unobservability of other minds" and was published in *Philosophical Explorations* in 2013. This article largely draws on conceptual analysis and philosophical clarification. Initially, I wrote the first two parts of the paper and my co-author wrote the third part. The manuscript then went through several rounds of rewriting by both of us, which was a source of great intellectual enjoyment because the text improved and condensed with every round. The idea for the paper was born during my stay at the Center for Subjectivity Research in Copenhagen, where I started to seriously think about the interactionist critique, which is based on the claim that theories of mindreading make the assumption that other minds are unobservable. Very simply put, interactionists argue as follows:

Premise 1: Theories of mindreading are based on the assumption that other minds are unobservable.
Premise 2: Other minds are not unobservable, because we can directly perceive (some) mental states of others.
Conclusion: Ergo, theories of mindreading are mistaken.

This kind of argument is ubiquitous in the critical accounts of interactionists: almost every interactionist account begins with a critique of the alleged underlying unobservability assumption of theories of mindreading. After systematic

re-reading of the standard accounts of mindreading, I became convinced that interactionists are attacking a straw man. The article explains why the interactionist critique is problematic. It analyses metaphysical, phenomenological, epistemological, and psychological readings of the unobservability claim and proves that it is not the case that theories of mindreading assume the metaphysical, phenomenological, or epistemological unobservability of other minds. However, theories of mindreading support a psychological version of the unobservability claim – a claim about cognitive processes responsible for mindreading. The psychological claim can be interpreted in a stronger or weaker sense. It can be read as a claim that "(a) neither the other's 'mindedness' in general nor the other's particular mental states are observable (i.e., apprehended perceptually); (b) particular mental states are unobservable, whereas some aspects indicative of 'mindedness' are observable; or (c) some mental states are unobservable but some are also observable" (Bohl & Gangopadhyay, 2013, p. 1). The critics tend to attribute the strongest psychological claim (a) to the proponents of theories of mindreading, but most of them actually support the weaker positions of (b) or (c). The conclusion of the article is that the allegations against theories of mindreading are seriously misdirected. The remainder of the article investigates constraints on the observability of other minds. It is argued that, given Husserl's phenomenological analysis of the structure of perception, any satisfactory account of social cognition must take into consideration the constraint that mental states and sensory properties of physical objects are not observable in the same way. It is shown that theories of mindreading are in a good position to deal with the above-mentioned stipulation and thus that interactionism has something to learn from theories of mindreading.

The third research article is entitled "We read minds to shape relationships." It was published in *Philosophical Psychology* in 2014. I consider it to be the most important part of my dissertation, not only because I am the sole author of the article, but because it makes an important original contribution by introducing a new dimension into the research on social cognition: the dimension of *social relationships*. Social relationships are extremely important for humans and our social behaviour largely depends on them, but the role of social relationships has been ignored by both theories of mindreading and interactionism. I introduce a theory of basic structures of social relationships and their cognitive underpinnings – *the relational models theory* by Alan Fiske – and use it for thinking about the function of explicit mindreading in human social cognition. As I will explain in chapter 2 of the summary article, a lot of effort has been made to develop theories of the cognitive mechanisms for mindreading. However, the issue of the *function* of mindreading in human social cognition has remained uninvestigated because of the prevalent assumption that mindreading is ubiquitous in human social cognition and serves the purpose of explaining and predicting other people's behaviour. I argue that the standard view does not capture the specific role of mindreading. Working from the relational models theory of Alan Fiske, I outline a hypothesis that the

evolutionary function, as well as the individual purpose, of mindreading is to monitor and shape social relations. I cache out the hypothesis into empirically testable claims and I suggest *experience sampling* as a possible method for testing them. I put forward reasons in favour of the view that relational cognition is more fundamental than mindreading in human social cognition. I pay separate attention to the issue of the motivational mechanisms of mindreading and hypothesise that explicit mindreading is often motivated by social emotions.

Although, the three research articles target the empirical debate about human social cognition, first and foremost they make a philosophical contribution. What is the role of philosophy in the interdisciplinary research of social cognition? The following chapter focuses on this issue.

1. THE ROLE OF PHILOSOPHY IN SOCIAL COGNITION RESEARCH[4]

Among other things, philosophy invites us to understand understanding.

Alvin Goldman

Like many other topics of cognitive science, questions related to *social* cognition have first been posed and addressed by philosophers. For instance, Descartes asked how we come to apprehend other people as having minds when we only see their bodies. He concluded that since we do not directly *see* other people's minds, we need to *infer* that they have minds by using the faculty of judgement:

> If I look out of the window and see men crossing the square, as I have just done, I say that I see the men themselves, just as I say that I see the wax; yet do I see any more than hats and coats that could conceal robots? I judge that they are men. Something that I thought I saw with my eyes, therefore, was really grasped solely by my mind's faculty of judgment. (Descartes, 1998, p. 68).

The classical philosophical problem of other minds is the question "How can we justify the belief that there are other minds besides our own?" The question at the centre of the current interdisciplinary research on social cognition is related but different: "What cognitive processes enable and sustain people's understanding of, and interaction with, others?" This question is also at the focus of the current doctoral thesis. It is first and foremost an empirical question, so one may wonder: Is philosophy at all necessary for answering this question, and if so, how? In this chapter, I will take a closer look at the relevance of philosophy to interdisciplinary research on social cognition. I will also bring out how I have used philosophical methods in the research articles that form the main part of my doctoral thesis.

It has been said that philosophy is to cognitive science as tin cans tied to a car are to a wedding or as alcohol is to sex (Thagard, 2009, p. 237). These comparisons express a serious doubt about what philosophical methods can contribute to the answering of empirical questions about cognition. One might be equally sceptical about the role of philosophy in the study of *social* cognition. Even when there is a sense that philosophy should be part of interdisciplinary research on social cognition, there is a lack of clear understanding about what philosophers *can* and *should* contribute to this field and how to effectively combine empirical and philosophical methods. When I started doing

[4] This chapter is based on Bohl, V. (2011). Milleks on sotsiaalse tunnetuse uurimisvaldkonnas tarvis filosoofiat? (Why does social cognition research need philosophy?) *Studia Philosophica Estonica* 4.1, 20–51.

research on social cognition, I soon came to ask myself: what it is that I *as a philosopher* can contribute to this field? Does my research make me less of a philosopher and more of a scientist? What are the distinctive philosophical methods that could be used to help solve puzzles about human social cognition?

Let me begin by borrowing the distinction between philosophy *in* and philosophy *of* cognitive science from Andrew Brook (2009). When a philosopher's focus of study coincides with the object of scientific research, for example when a philosopher asks "How does social cognition work?", we are dealing with philosophy *in* science. In contrast, when a philosopher makes scientific research itself an object of inquiry by asking, for example, "What is good science?" or "What are scientists doing when they are studying social cognition?", we are concerned with philosophy *of* science. Since the role of philosophy of science is relatively well understood and appreciated (Brook, 2009, p. 218), and since it is not the focus of my thesis, I will here primarily analyse the second issue: What is the contribution of philosophy *within* research on social cognition?

We can distinguish between two questions: "Do we need *philosophy* in research on social cognition?" and "Do we need *philosophers* in research on social cognition?" I think that when we answer "yes" to the first question, the second question also needs to be answered in the affirmative and vice versa. One might want to argue that instead of hiring professional philosophers, scientists should be trained in philosophical methods, so that scientists would be able to handle the philosophical part of research. But this would mean that scientists would partly *become* philosophers, so it would not change the claim that if philosophy is needed, so are philosophers. Both of these questions converge at the issue of whether philosophical *competence* can contribute to research on social cognition. In the following, I will argue that it can and it should.

1.1. Philosophy is not just "armchair speculation"

There is a widespread view that, although philosophers have in the past come up with many interesting speculations on important (e.g. mind-related) issues, today various scientific disciplines are in a position to provide empirically informed answers to most of the questions that long ago used to belong to philosophy (Brook, 2009, p. 219). Therefore, it is argued, philosophy has lost its relevance in many areas. For example, philosophers are no longer working on questions such as "What is life?" (at least not the way they used to); this issue now belongs to modern biology. One could argue that social cognition is another such area wherein philosophy has no real role to play.

It is important to notice that the view above implies that when philosophers are studying social cognition, they are in principle not doing anything different from what psychologists, neuroscientists, or other researchers are doing – they are looking for answers to the very same questions that scientists are wrestling with. The problem is, it is argued, that philosophers are not doing it very well, since they lack the necessary empirical tools and they thus merely engage in a

poor substitute for proper empirical research – armchair speculations. If this is the case then philosophers are simply doing bad science.

These allegations strike me as doing an injustice to philosophy. Indeed, except for experimental philosophers, philosophers are not collecting data or running experiments. But although philosophy is not empirical research in the narrow sense of data collection, philosophers are in a good position to develop *theories* and to generate new empirically testable *hypotheses*. While theorising and hypothesising, philosophers are not necessarily doing anything principally different from what scientists are doing when they theorise or come up with hypotheses. However, philosophers can have a distinctive role in this process. This is because: they are especially well trained in critical thinking; they have (or should have) background knowledge of the history of philosophical thought; and their theoretical interests are typically somewhat broader than those of most empirical scientists.

The field of social cognition is rich in examples of fruitful collaboration between philosophers and empirical scientists where theoretical claims and hypotheses first offered by philosophers have subsequently been experimentally tested and the results of the experiments have in turn been integrated into theoretical discussions. For example, the classic paper "Does the chimpanzee have a theory of mind?" (by primatologists Guy Premack and David Woodruff (1978)) is a report of experiments that were inspired by the ideas of Daniel Dennett (see Dennett, 2009, p. 233) and influenced by the functionalist conception of mind (see Morton, 2009, p. 714, Goldman, 2006, p. 10), as well as by prominent ideas in philosophy of science (Goldman, 2006, p. 11). Later on, in commentaries to the paper, three philosophers – Dennett (1978), Bennett (1978), and Harman (1978) – independently argued that the possession of the concept of *belief* is crucial to understanding others' mental states. They came up with the idea that one could test whether a person possesses the concept of belief by investigating the ability to attribute *false* beliefs. As a result, psychologists Heinz Wimmer and Josef Perner created the first *false belief task* in 1983, which became an important landmark in the developmental psychology of social cognition – it triggered extensive research on children's development of meta-representational abilities.

Another important role of philosophers besides theory and hypothesis building is to critically evaluate the theoretical frameworks and methodologies of current research and to suggest new research directions. For example, philosophically inspired critique of the so-called third-person mindreading paradigm (see e.g. Leudar & Costall, 2009a) has fostered the search for new methods that would enable the investigation of more direct second-person interactions between people in more natural settings (Schilbach et al., 2013). New methodological frameworks lead to new discoveries; for example Wilms et al. (2010) have recently shown that when a person is interacting with someone from a second-person perspective, there are significant differences in her brain processes as compared with a situation in which she is merely observing another subject from a third-person perspective.

What is the relationship between philosophical reflection and empirical findings? As far as social cognition is concerned, although philosophers are not collecting empirical data, they frequently rely on empirical evidence to support their theories. Thus they are not to be blamed for succumbing to *a priori* armchair speculations *instead* of finding out how things really work in the world. For example, in a recent book "Simulating Minds", Alvin Goldman (2006) relies extensively on neuroscientific and psychological evidence to support his version of simulation theory. Nevertheless, no piece of evidence by itself proves a theoretical claim – it first needs to be interpreted accordingly. For example, Goldman refers to evidence of mirror neurons in humans in order to argue for the simulation theory. However, this evidence also allows for different interpretations so that it can serve as support for a non-simulationist approach (compare e.g. Gallese & Goldman, 1998, Gallagher, 2007, and Slors, 2010 for different accounts of the function of mirror neurons). This demonstrates that theoretical issues cannot be solved by simply piling up more empirical results. Rather, conceptual clarification and a scrutiny of assumptions that are implicit in the current research paradigm are often necessary. For this kind of work, philosophers are suitably "armed." At the same time, without new empirical findings, there would be little progress in theories of social cognition. So it is best to see empirical data collection and theory-building (including the use of philosophical analysis) as complementary. It is not a coincidence that since empirical disciplines and philosophy joined forces to study social cognition, more elaborate accounts of social cognition have emerged than there ever did in the course of the rest of the history of philosophy and science.

These and other similar examples of successful cooperation between philosophy and the empirical sciences demonstrate that philosophical discussions have guided empirical research of social cognition in important ways and I see no reason why this should not continue. Obviously, successful interdisciplinary cooperation demands a special effort from both philosophers and scientists. A meaningful exchange is not always easy to achieve. An expected problem is that theoretical claims originating within philosophy tend to be too abstract or too general to be linked with particular empirical findings or to allow for direct empirical testing. For example, Ian Apperly (2011, p. 5) charges that "nobody has yet come up with a generalizable test, or set of criteria, that could be used to discern whether a given mindreading problem was solved by simulation or by theorizing" and argues that for this reason the philosophical debate between the proponents of the two major theories of how people attribute mental states – theory-theory and simulation theory – is not particularly useful from the point of view of experimental psychology. This highlights the need to focus more on the operationalization of philosophical concepts without distorting or trivializing them. But it also shows that in frontier science, figuring out what questions to ask and how to ask them is not a trivial matter. As Dennett (2009, p. 232) has put it:

> One of the reasons cognitive science is such a land of plenty for philosophers is that so many of its questions – not just the grand bird's-eye view questions but quite proximal, in-the-lab-now questions – are still ill thought out, prematurely precipitated into forms that deserve critical reevaluation. If philosophy is, as my bumper sticker slogan has it, what you're doing until you figure out just what questions to ask, then there is a lot of philosophy to be done by cognitive scientists these days.

There is, no doubt, also a lot of philosophy to be done by researchers of social cognition.

1.2. Painting with broad brush strokes?

Another distrustful view on what philosophers are doing in science is that unlike scientists, philosophers are only interested in the most general questions and in the "big picture" of the world. They thus tend to paint with brush strokes that are too broad (compare Brook, 2009, p. 219). Of course, science is also interested in questions of a general nature but according to the critics, philosophical ambitions go far beyond empirical evidence and bring along a risk of ignoring, oversimplifying, or distorting scientific facts.

It is probably true that "philosophers often comically misjudge their competence to evaluate concepts, arguments, theories with which they have only the most passing acquaintance" (Dennett, 2009, p. 232). However, the development of more general theoretical frameworks is certainly necessary when it comes to issues that call for an interdisciplinary approach. Social cognition is one of the areas of research that does not fit nicely within the boundaries of only one discipline – the study of social cognition has become a truly interdisciplinary enterprise involving philosophy, psychology, cognitive science, neuroscience, and various other disciplines. However, interdisciplinarity brings along its own problems. On the one hand, people who do interdisciplinary work must be experts in their own field of research and know where the limits of their competence and the boundaries between disciplines lie: strong fences make good neighbours. On the other hand, truly interdisciplinary work requires that specialists are able to overcome the theoretical, conceptual, and methodological limits of their home discipline, see the object of study from a wider perspective, and pose questions from a more general point of view. But who should be reflecting upon how the theories and empirical results of different disciplines hang together? Who should clarify questions that surge up at the intersection of different disciplines? Philosophers are good candidates for this task because they have both the means and the motivation to think about how results of various disciplines hang together (or contradict one another) on a larger scale. In addition to fulfilling the general desire for a more coherent and holistic understanding of the world, such an integration process potentially leads to new empirical questions, stimulates the development of new methods, and thereby creates new knowledge. Research results of one discipline can set constraints on or provoke questions about processes that could be better investigated by using

the means of some other discipline. In short: philosophers can and should have a distinctive role in interdisciplinary research (see also Sellars, 1962, Thagard, 2009, p. 238).

The articles that form the main body of my thesis have largely stemmed from a truly interdisciplinary approach to issues related to social cognition. In appendix 1, with the neuroscientist Wouter van den Bos, I integrate aspects of the "theory of mind" paradigm with aspects of phenomenologically inspired interactionist approaches by relying on dual-process theories, to build a new comprehensive theoretical framework. The process of developing an integrative framework also gave rise to a new empirical hypothesis: the idea that in social interaction, Type 1 and Type 2 processes mutually influence each other. In appendix 3, I integrate the perspectives supplied by cognitive and developmental psychology on mind-reading with the cognitive anthropology of relational cognition in order to argue that the function of mindreading is to shape social relationships. The main claim of the paper is developed into empirical hypotheses and supplemented by a description of a possible method for testing the hypotheses.

Not all attempts to integrate the findings of different disciplines with philosophical insights result in new empirical hypotheses. Sometimes the result is more abstract. Whether theories that have gained a certain level of generality and abstractness can be called empirical science is debatable but the gap between general claims and empirically proven particular claims is in principle similar to the problem of induction – a problem that is unavoidable in most scientific thinking. Philosophers engaged in interdisciplinary research contribute to the answering of empirical questions but they also sometimes weigh empirical findings in order to answer specifically philosophical questions. Philosophically oriented theories that rely on scientific research for finding answers to traditional philosophical problems can be called *empirical philosophy* instead of empirical science (Prinz, 2008). The boundary between empirical science and empirical philosophy is, however, not sharp. We can think of traditional philosophy and data collection rather as two ends of a continuum (Prinz, 2008, p. 206). When it comes to the danger that philosophers may easily misunderstand or overlook scientific facts, it is in every philosopher's interest to get the facts straight, since any theoretical claim that ignores or misrepresents facts becomes an easy target for its critics. This ensures the interest of naturalistically minded philosophers[5] in keeping up with empirical research that is

[5] With the expression "naturalistically minded philosophers" I have in mind philosophers who may not agree upon the definition of naturalism but who, one way or another, think that philosophy needs to take scientific knowledge seriously. It includes philosophers who support either a) ontological naturalism (the claim that everything that exists is natural, i.e. no supernatural entities or processes exist); b) methodological naturalism (a claim that we can aquire true knowledge about the world only by using the methods of natural science or a claim that philosophy and science use the same kinds of methods in principle (see Papineau, 2009a)); c) a broad definition of naturalism (see e.g. Zahavi, 2004) which emphasizes that philosophers must take scientific knowledge into consideration without necessarily agreeing to ontological or methodological naturalism.

relevant to their field of interest. At the same time, it is also a philosopher's job to pay close attention to how scientific results are produced and to scrutinize the tacit assumptions that are left implicit in research routines.

1.3. Conceptual hygiene

Analytic philosophy is known for its emphasis on conceptual analysis, so one way to fit philosophy into empirical research is to give philosophers the job of taking care of the conceptual tool kit of the research field. Empirical scientists, however, may not like the idea of philosophers telling them which concepts they should be using and how. Also philosophers themselves have criticised the view that conceptual analysis should be an important part of philosophy. For example, Paul Thagard (2009) claims that philosophy operates best, not in the form of conceptual analysis, but rather as empirically informed reflection on a wide range of scientific findings. Jesse Prinz (2008) does not deny the necessity of conceptual analysis but argues against the view that it is a task for philosophers in particular. According to David Papineau (2009b), philosophers who think that they are doing conceptual analysis are simply confused about what it is that they are actually doing.

Conceptual clarity is certainly important not only in philosophy but also in science. Unfortunately, the conceptual tool kit of social cognition research is not in very good shape. Theorists of different backgrounds are frequently using the same terms for different concepts and different terms for the same concepts, which causes a great deal of confusion. It often leads to talking past one another and creates merely verbal disagreements. For example, in the framework of theory-theory and simulation theory, "social cognition" is primarily used for the ability to predict and explain other people's behaviour by attributing mental states to them. Anti-mindreading approaches, such as interactionism (see chapter 3), use "social cognition" primarily to refer to the ability to successfully coordinate behaviour in social interactions. Another example is that the term "theory of mind" is sometimes used as a synonym for the ability to mindread (i.e. the ability to attribute mental states), sometimes for the theory-theory of mindreading, and sometimes as an umbrella term for a whole set of mainstream theories of mindreading (including both theory-theory and simulation theory, which I introduce in the next chapter). It is obvious that in order to avoid confusion and merely verbal disagreements, different parties should adjust their concept use or at least be aware of the fact that their use of some terms differs from that found in other contexts.

There are various ways to improve conceptual clarity. The usefulness of conceptual analysis depends on the exact method and the aims of the analysis. According to Tim van Gelder (1998, pp. 122–124), philosophers have at least three standard methods for conceptual analysis: 1) they consult their personal intuitions "from an armchair"; 2) they analyse how particular words and phrases are used in certain contexts (van Gelder labels this method "linguistic

analysis"); 3) they work out new concepts. A fourth method, known as *experimental philosophy*, could be added: 4) the empirical study of how ordinary people use the concepts of interest.

Often, philosophical conceptual analysis is identified with the first, "armchair," method. This method implies that philosophers at least implicitly possess the concepts in question and that, by analysing one's intuitions, it is possible to make them explicit: "armchair conceptual analysis can be characterized as an introspective memory retrieval process" (Prinz, 2008, p. 191). The problem with this method is that, without due reason, it implies that philosophers' intuitions are more trustworthy than the intuitions of non-philosophers. It is questionable whether philosophers' intuitions about vernacular concepts are more reliable than those of non-philosophers. Even if it is the case, an additional complication is that the intuitions of different philosophers often contradict each other, so it is unlikely that any universally valid intuitions about concepts exist. In any case, it is highly implausible that philosophers' intuitions about scientific concepts are more reliable than those of scientists.

The second method is more likely to be useful for keeping the conceptual tool kit of a research field in good shape. Here the philosopher's job is to apply her trained analytical skills to figure out how key terms are used in different research contexts. Based on such an analysis, it is possible to disentangle different concepts and to make suggestions about how particular terms could be used more clearly to avoid talking past one another and to prevent or resolve verbal disagreements. This kind of analysis of concepts is certainly of great value for research, especially because most scientists (except for linguists, but their motivation is different from that of philosophers) lack the time and skills to carry it out.

Both of these methods assume that the concepts of interest already exist "ready-made" either implicitly in one's mind or within the linguistic practice of a particular (scientific) community. For this reason, conceptual analysis is seen as a process of making concepts explicit and linking them with specific terms (van Gelder, 1998, p. 123). However, such an assumption cannot be taken for granted. It is possible that everybody is confused about what is, for example, "social cognition" or "theory of mind." If this is the case then the main task of conceptual analysis is not to make implicit concepts explicit but rather to *create* new and more effective concepts. The conceptual work becomes a prescriptive, rather than a descriptive, process. Whether any particular newly constructed concept will be adopted by a wider range of researchers, only time will tell.

We have yet to discuss the fourth method – the one adopted by experimental philosophers. Experimental philosophers run questionnaire-based studies to uncover the intuitions of "ordinary folk" concerning concepts of central philosophical importance (Knobe & Nichols, 2008). In some contexts, for example in ethics or epistemology, this method can provide important philosophical insights. Some experimental philosophers even argue that this method can be used to determine whether the intuitions of armchair philosophers are correct (Prinz, 2008, p. 201). However, in the context of social cognition research (as is

likely in many other scientific contexts), consulting vernacular concepts is unlikely to be helpful in building the conceptual apparatus of the field. The "ordinary folk" simply does not possess fine-grained concepts for describing and explaining the cognitive processes that are responsible for human social cognitive abilities.[6]

We can conclude that just as with any other field of research, conceptual "hygiene" is important for research into social cognition and that philosophers are professionally well equipped for this job, especially when they take up linguistic analysis and work out new "tailor-made" concepts.

The application of the tools of conceptual analysis is often implicit in any philosophical writing but I have more explicitly relied on conceptual analysis in appendix 2, where I and my co-author Nivedita Gangopadhyay analyse how the concept "unobservable" is used by proponents and critics of theories of mindreading. We disentangle four different senses of "unobservable" and show that confusing them has caused a great deal of verbal disagreement between the proponents and the critics of theories of mindreading. By clarifying the different meanings of the concept, we dissolve elements of verbal disagreement and show that the real issue of controversy between the theories of mindreading and the critics of it lies in the question of how to explain the cognitive processes that underlie mental state attributions.

1.4. The role of phenomenology in social cognition research

Until now, I have focused mainly on the role of traditional and analytic philosophy in the research on social cognition. In recent years, philosophers with a background in phenomenology have started to enter the interdisciplinary debates concerning social cognition. Phenomenology is a philosophical discipline founded by Edmund Husserl at the beginning of the 20th century which focuses on the study of the structure of experience. How can phenomenology contribute to research on social cognition?

An obvious answer is that just as with many other philosophical works, the classical texts of Edmund Husserl, Martin Heidegger, Maurice Merleau-Ponty, Jean-Paul Sartre, Emmanuel Levinas, Hannah Arendt and other phenomeno-

[6] However, experimental philosophy can be useful for studying what in philosophy goes by the name "folk psychology" (according to one prominent definition of "folk psychology"): principles concerning human psychology that ordinary people are inclined to endorse in their interpretations of everyday human behaviour. Philosophers working on folk psychology tend to prematurely assume that they know how "the ordinary folk" reasons about the mind. For instance, it was thought that according to folk psychology, "intentions" are mental states that precede deliberate actions. By using the methods of experimental philosophy, Knobe (2008) showed that folk psychological attributions of intentions are much more complex: whether an act is judged to have been carried out intentionally or not depends on the moral consequences of the act.

logists can serve as a reservoir of theoretical insights that can potentially be used as a basis for working out scientific hypotheses. However, for Husserl, phenomenological descriptions were supposed to operate on a more fundamental level than any scientific claim. He considered phenomenological truths to be untranslatable into scientific language. From a Husserlian point of view, phenomenology is a transcendental discipline for investigating the *a priori* structures of experience, which are *presupposed* by any scientific practice. Phenomenologists would therefore not necessarily welcome attempts to naturalize phenomenology or to make phenomenology continuous with science (see Zahavi, 2010, p. 7). Nonetheless, some great figures of the phenomenological tradition have participated in a fruitful dialogue with the science of their time. One of the best examples of this is the great French phenomenologist Maurice Merleau-Ponty whose "Phenomenology of Perception" (Merleau-Ponty, 2002, originally published in French in 1945) is informed by psychological research of the time and who famously analysed and reinterpreted various psychopathological findings within a phenomenological framework. Thus it is an open question whether phenomenology can and should be naturalized. Although there is a lack of consensus among phenomenologists about this issue, one option for reconciling phenomenology with naturalism is to understand naturalism in a broad enough sense so that within the naturalistic framework the *subjectivity* of experience is done justice. "[W]hy let the reductionists monopolize the concept of naturalism?" asks Zahavi (2004, pp. 343–344), proposing that naturalism needs to be redefined so as to encompass phenomenology without reducing phenomenological claims to scientific claims. In other words, one could understand naturalized phenomenology as that part of phenomenology which allows for a mutual exchange of ideas between science and phenomenology (Zahavi, 2010, pp. 14–15).

The use of phenomenology in the interdisciplinary study of social cognition is, unfortunately, hindered by a widespread view according to which phenomenology is restricted to describing experiences from the subject's personal point of view. The standard accounts of social cognition (theory-theory and simulation theory) treat phenomenological descriptions as irrelevant to research on social cognition, since it is thought that the descriptions of subjective experiences do not provide any useful information about sub-personal social cognitive mechanisms. However, it should be pointed out that non-phenomenologists often misunderstand the nature of the phenomenological methodology and wrongly identify it with introspection (see Gallagher & Zahavi, 2008, pp. 13–43). Since introspection as a research method is considered to be unreliable and unsuitable for scientific research, it is no wonder that scientists and many philosophers are sceptical about the prospects of phenomenology contributing to the solution of puzzles raised by social cognition. It may come as a surprise that phenomenologists are actually even more critical of the concept of introspection than scientists: they do not just claim that introspection as a method is unreliable but argue that introspection is a confused notion that corresponds to no real process. According to phenomenologists, the concept of introspection

draws upon a false distinction between the "inner" and the "outer"; it assumes that consciousness is locked inside one's head and the world is outside of it (Gallagher & Zahavi, 2008, p. 21). Phenomenologists argue that subjective experiences are not locked inside our heads. They are ways in which we are open to the world as embodied beings. One of the aims of phenomenology is to show that experience and consciousness cannot be reduced to or fully explained from a scientific third-person perspective. Merleau-Ponty (2002, p. ix) expresses it as follows:

> Scientific points of view, according to which my existence is a moment of the world's, are always both naïve and at the same time dishonest, because they take for granted, without explicitly mentioning it, the other point of view, namely that of consciousness, through which from the outset a world forms itself round me and begins to exist for me.

Scientists and many naturalistically oriented philosophers tend to see science as fundamental for providing knowledge about the world, but traditional Husserlian phenomenologists see phenomenology as more fundamental because it is seen as providing a foundation for science. I will not try to solve this contradiction here but I think that a fruitful cooperation between science and phenomenology is possible even without agreeing on a solution to this problem. For instance, there is no reason why phenomenological descriptions could not be treated as *explananda* for social cognition research: phenomenology could provide descriptions of first-person experiences that could at least partly be explained by the underlying cognitive mechanisms. Authors who criticise the standard accounts of social cognition appeal often to the argument that the *explanandum* of these accounts is phenomenologically inadequate (see e.g. Leudar & Costall, 2009b). But there is a further issue of whether providing *explananda* is the only thing phenomenology can contribute. Can phenomenology provide any explanations?

Shaun Gallagher (2003) has written about three different ways in which phenomenology can contribute to experimental science. The first option is *neurophenomenology*, as originally outlined by Francisco Varela (1996) and applied in pilot experiments by Lutz et al. (2001; see also Lutz, 2002). The idea of neurophenomenology is that phenomenologically exact first-person data informs the interpretations of physiological processes and, in turn, neuroscientific analyses help to refine phenomenological descriptions. In standard neuroscientific experiments, individual findings are highly variable but the variability is usually attributed to various subjective parameters that are treated as noise. The "noise" is "washed out" by averaging results across subjects and different trials. The neurophenomenological approach, however, treats first person data as a valuable source of information and combines it with a dynamical analysis of neural processes. The phenomenological part of this approach involves using phenomenological methodology to describe subjective parameters through a series of trials. Both empirical scientists and experimental subjects need to be extensively trained in phenomenological methods, in particular to be able to

shift their attention from *what* they experience to *how* they experience it in order to deliver consistent and clear reports of the experience. The subject's reports are used as descriptive categories to divide the trials into phenomenologically based clusters, which are then correlated with dynamic neural signatures. (see Gallagher, 2003, pp. 86–88).

The second option that Gallagher (2003, pp. 88–91) describes is what he calls *retrospective and indirect use of phenomenology.* It involves using phenomenological insights to critically reinterpret theoretical claims and experimental results much in the way Merleau-Ponty used phenomenology to re-evaluate the scientific claims of his time. However, this approach is incomplete unless the phenomenological reinterpretations are in turn empirically tested.

This brings me to the third option, which Gallagher labels *front-loaded phenomenology* (ibid., pp. 91–97): using phenomenological insights to shape the way experiments are designed. More generally, the method should lead to a dialectical movement between phenomenological insights and preliminary trials, a movement which is aimed at specifying claims and concepts for the purposes of the experiments. Gallagher gives as an example the phenomenological distinction between *self-agency* and *self-ownership* that has informed several neuroscientific experiments, which have themselves helped refine the phenomenological distinction (see Farrer & Frith, 2002, Chaminade & Decety, 2002). There are also recent examples of front-loading phenomenology to social cognitive neuroscience. Schilbach et al. (2006) have used fMRI to demonstrate that the phenomenological distinction between the second- and the third-person points of view (between the experience of being involved in social interaction and the experience of being a passive observer of social interaction) is correlated with different neural patterns in the two conditions. The distinction has inspired a whole new research paradigm in social cognitive neuroscience, outlined in a recent article by Schilbach et al. (2013).

When we look at what role phenomenology has played in current research on social cognition, it just happens that it has mainly been used indirectly and retrospectively. To my knowledge, there have been no neurophenomenological experiments targeting social cognition and the contribution of front-loaded phenomenology is relatively modest. The main strategy for including phenomenology within the study of social cognition has been to use phenomenological or phenomenologically inspired arguments to criticise current theories of social cognition and to reinterpret empirical findings so that they would be more in accordance with phenomenological insights. For example, based on the views of Scheler, Sartre and other classical phenomenologists, Gallagher and Zahavi (2008) argue that instead of mindreading via theorising, social cognition is primarily a specific perceptual or empathic process targeted at the living body of the other person. They also reinterpret the function of mirror neurons in light of the phenomenologically inspired direct perception approach and thereby also oppose the simulationist interpretations of social cognitive processes.

As for the role of phenomenology in my thesis, in the appendix 2, I and my co-author Nivedita Gangopadhyay use Husserl's account of the structure of

social perception to argue that, ironically, phenomenologically inspired critics of the theory of mind approach tend to oversimplify the phenomenon of social cognition by characterising it as similar to non-social perception. We argue that Husserl's account is not necessarily incompatible with theory of mind accounts, and that Husserl's insights could help to develop a more adequate framework for studying social perception. In Husserl's phenomenological analysis, the structure of social perception differs in important ways from the structure of perception of non-social entities: mental states are not perceivable in the same way as sensory properties of objects. In the future, it will be interesting to try to find ways to empirically test these claims (or more specific hypotheses based on them) by front-loading phenomenological insights into the experimental design.

1.5. Conclusion

In this chapter, I have focused on the issue of how philosophical methods can be used to study *social cognition*. Philosophers are not necessarily doing anything categorically different from that which scientists do when they try to answer questions such as "How does social cognition work?" However, because they possess a background knowledge that encompasses the history of thought, special training in critical thinking, and motivation to unravel questions that exceed the boundaries of established scientific disciplines, philosophers can play a distinctive role in research on social cognition. Philosophers are in a good position to critically evaluate the theoretical frameworks and methodologies of ongoing research projects. Philosophical competence conjoined with a drive to understand the human mind and sociality provide a valuable basis for a truly interdisciplinary approach to social cognition and fosters efforts to build more comprehensive theoretical frameworks to cover various aspects of human social cognition. Philosophers have the necessary skills to reflect upon how the theoretical claims and empirical results of different disciplines and paradigms hang together (or contradict one another) on a wider scale. Integration of philosophical insights and research results stemming from different disciplines opens up new research avenues, raises important questions at the intersection of disciplines, stimulates theoretical discussions, and contributes to the development of new hypotheses and research methods. In addition, because philosophers are trained in conceptual analysis, they are well suited to do the necessary conceptual work. Philosophers can best take care of the conceptual tool kit of the field, not by consulting their personal intuitions on particular concepts, but rather by analysing existing conceptual frameworks and by "tailor-making" new concepts. As far as phenomenology is concerned, it can contribute to the research of social cognition in several ways: 1) phenomenological descriptions of social experiences can serve as *explananda* for social cognition research; 2) by teaching the phenomenological method to test subjects, it is possible to carry out neurophenomenological experiments; 3) the phenomenological framework

can be used retrospectively and indirectly to reinterpret scientific findings; 4) phenomenological insights can be "front-loaded" into the experimental design.

The research articles that constitute the main part of my doctoral thesis (appendices 1–3) include several ways of using philosophy *in* science as described above. In papers 1 and 3, contrasting theoretical and empirical perspectives for explaining the underpinnings of human social cognition are synthesised. This enabled me to outline more comprehensive theoretical frameworks but also resulted in novel empirical hypotheses and suggestions for testing them. The integrative approach outlined in "Toward an integrative account of social cognition: Marrying theory of mind and interactionism to study the interplay of Type 1 and Type 2 processes" (co-authored with Wouter van den Bos, see appendix 1) merges the so-called theory of mind approach and the interactionist approach with the help of dual process theories. It also presents the hypothesis that the two types of cognitive process (Type 1 and Type 2 processes) that have so far been studied in isolation, and that have been thought to function independently, may interact and mutually influence each other in real life social interactions. In "We read minds to shape relationships" (see appendix 3), I focus upon an issue that has received little attention in the mainstream social cognition literature – the function of mindreading. By introducing the relational model's theory of Alan Fiske, it is hypothesised that the primary function of mindreading is to shape (create, sustain, negotiate etc.) social relationships. The hypothesis is cashed out in the form of specific empirical claims and experience sampling is suggested as a possible method to test those claims. The article "Theory of mind and the unobservability of other minds" (co-authored with Nivedita Gangopadhyay, see appendix 2), is mainly aimed at conceptual clarification. It is shown that "unobservability of mental states" can be read in various ways. Most of the readings, contrary to what the critics of theory of mind have assumed, do not apply to theories of mindreading. This leads to the conclusion that the disagreement between the proponents and the critics of theories of mindreading over whether mental states are observable is largely a red herring. The second part of the paper makes use of Husserl's analysis of the structure of social perception, thus offering an example of how phenomenological insights can enter into the study of social cognitive processes.

2. THE STANDARD APPROACH TO SOCIAL COGNITION: READING EACH-OTHER'S MINDS

Attribution of mental states is to humans as echolocation is to the bat.

Dan Sperber[7]

Having introduced the meta-philosophical and methodological issues pertaining to the role of philosophy in the study of social cognition, we can now move on to the topic of social cognition itself. In this chapter, I give an overview of how social cognition has been studied within the past three to four decades by researchers who have assumed that the key to human social cognition is *mind-reading* – the ability to attribute mental states in order to make sense of other people's behaviour. I review the most influential experimental studies of the standard mindreading paradigm and introduce the mainstream theories of mind-reading: *theory-theory* and *simulation theory*.

2.1. From Premack and Woodruff's chimpanzee experiments to the false belief tasks

In 1978, in a special issue of *Behavioural and Brain Sciences* on consciousness and cognition in non-human species, primatologists David Premack and Guy Woodruff published a seminal article "Does the chimpanzee have a theory of mind?" The article provoked the interest of philosophers and psychologists and triggered interdisciplinary research on the ability to understand mental states. The authors assumed that humans possess a "theory of mind" – an ability to impute mental states to oneself and others via theoretical inferences. They were interested in whether chimpanzees also have something like a theory of mind. On their account, the ability to impute mental states has a theoretical basis because "such states are not directly observable, and the system can be used to make predictions about the behaviour of others" (Premack & Woodruff, 1978, p. 515). Although the paper by Premack and Woodruff lacks references to philosophical works, it was clearly influenced by the contemporary functionalist conception of mind[8] (see Morton, 2009, p. 714, Goldman, 2006, p. 10), and the accounts of "theoreticity" in the philosophy of science (see Goldman, 2006, p.

[7] The quote from Dan Sperber originates from a paper presented at a conference on Darwin and the Human Sciences, London School of Economics, June 1993. Cited via Baron-Cohen (1997, p. 4).

[8] According to functionalism, mental states are functional states that figure in causal explanations of behaviour.

11). The paper was also inspired by an exchange of ideas between Guy Premack and Daniel Dennett (see Dennett, 2009, p. 233).

"Does the chimpanzee have a theory of mind?" is a report of a series of experiments with a chimpanzee called Sarah. In the first experiment, Sarah was shown videotaped scenes of a human actor struggling to obtain bananas that were in different unreachable locations. With each video she was given a pair of photographs of which only one represented a solution to the problem. Sarah consistently chose the "correct" photographs. The authors argued that by choosing the correct photograph Sarah expressed her understanding of the actor's *intentions*. Other possible interpretations were also considered (simple physical matching, associationism, and empathy or putting oneself in the place of the other) but the "theory of mind" explanation was preferred. The paper also reports several subsequent experiments that were designed to test different aspects of the "theory of mind" interpretation against other possible explanations. Premack and Woodruff concluded that chimpanzees may possess a primitive theory of mind.

The paper evoked a lively philosophical discussion in the commentaries. Philosophers Jonathan Bennett (1978), Daniel Dennett (1978), Gilbert Harman (1978), and the cognitive scientist Zenon Pylyshyn (1978) independently emphasized that in order to have a "theory of mind", an individual must be able not only to *have* beliefs but also to *represent* them: one needs to have the ability to have beliefs *about* beliefs. Several authors suggested that it is possible to test the existence of such a metarepresentational ability by examining an individual's ability to understand that people can represent the world differently from the way it really is. For example, Dennett (1978, p. 569) wrote:

> Very young children watching a Punch and Judy show squeal in anticipatory delight as Punch prepares to throw the box over the cliff. Why? Because they know Punch thinks Judy is still in the box. They know better; they saw Judy escape while Punch's back was turned. We take the children's excitement as overwhelmingly good evidence that they understand the situation – they understand that Punch is acting on a mistaken belief (although they are not sophisticated enough to put it that way). Would chimpanzees exhibit similar excitement if presented with a similar bit of play acting (in a drama that spoke directly to their "interests")?

Harman (1978, pp. 576–577) outlined a very similar idea:

> Suppose that a subject chimpanzee sees a second chimpanzee watch a banana being placed into one of two opaque pots. The second chimpanzee is then distracted while the banana is removed from the first pot and placed in the second. If the subject chimpanzee expects the second chimpanzee to reach into the pot which originally contained the banana, that would seem to show that it has a conception of mere belief.

These ideas were taken up by Austrian psychologists Heinz Wimmer and Josef Perner who designed the very first *false-belief task* for children. They published a paper (Wimmer & Perner, 1983) which described four experiments with children aged between three and nine years. The first two experiments tested children's ability to represent other subject's false beliefs; the other two experiments examined the ability to construct deceitful utterances without representing false beliefs.

Let us look at the first experiment in detail. Children were presented with puppet sketches where a character (a boy called Maxi, or a little girl) placed an object (a chocolate bar or a book) into one location and then witnessed how, in the absence of the character, the object was transferred to a different location. Children were asked to indicate where the protagonist will look for the object after returning to the scene ("Belief question"). Subsequently, they were asked either: a) where the protagonist will say the object is when he (or she) wants another person to help him (or her) to obtain it ("Cooperative condition"); or b) where the protagonist will say the object is when he (or she) wants another person not to find it ("Competitive condition"). The correct answer to the "Belief question" is that the protagonist will search for the object in the first location. None of the 3–4 year old children answered the first question correctly but 57% of 4–6-year old and 86% of 6–9-year old children pointed correctly to the second location. Children at all ages who correctly ascribed to the protagonist a false belief were also able to construct an utterance for the protagonist which was deceitful or truthful in relation to the protagonist's belief. Importantly, the majority of children who failed to answer the "Belief question" correctly nonetheless remembered where the protagonist had left the object, so the cause of the incorrect answer was not a failure to recall where the protagonist had left the object. In the second experiment, further conditions were added to the original stories to explore why the youngest subjects failed to ascribe a false belief to the protagonist. One hypothesis was that smaller children fail to answer the "Belief question" because they do not reflect on the situation and reply impulsively. To test this hypothesis, a "Stop-and-think" condition was introduced, where children were explicitly instructed to reflect on the situation before answering the "Belief question." Since the "Stop-and-think" condition did not improve younger children's performance, the hypothesis was not supported. The overall conclusion of Wimmer and Perner (1983, p. 126) was that "the emergence of children's ability to understand another person's beliefs … is not a mere side effect of an increase in memory and central processing capacity." Instead they suggested that within the period of 4 to 6 years a novel cognitive skill for representing false beliefs emerges.

The false-belief task soon became a kind of a "litmus test" for the presence of a theory of mind (Barresi & Moore 1996, p. 118), so much so that a decade after the first false-belief experiments Gopnik, Slaugther and Meltzoff (1994, p. 157) warned against "a neurotic task fixation" in developmental psychology. After the study of Wimmer and Perner there was an explosion of research on the astonishing improvement between 3 and 5 years of age in children's perfor-

mance on various false-belief tasks. Even today, 30 years after the first false-belief experiment, the development of false-belief understanding continues to be at the focus of developmental psychology.

Various versions of the false belief task have been designed and used on different populations. Experimenters typically use a two locations scenario (e.g. Maxi's chocolate might be in the drawer or the cupboard), or an unexpected contents scenario (e.g. the candy box that usually contains candies now contains something else). Numerous studies have replicated the result that younger children, typically 3-year olds, fail the standard false belief task, whereas older children, typically 5-year olds, pass it. An interesting finding in the two locations condition is that younger children are found to make a specific *false-belief error*: instead of answering randomly they typically assert that the protagonist will look for the object at the location to which it was moved (Wellmann, Cross & Watson, 2001, p. 656). In the unexpected contents condition (e.g. Gopnik & Astington, 1988), children who see a box of a well-known brand of candies state that it will have candies inside; then they open the box and find that it contains pencils. They are asked what another person who has not observed the contents of the box will think is in it. Most five-year olds answer correctly, whereas 3-year olds typically say that another naïve person will think that the box contains pencils. Interestingly, when the younger children are asked what they themselves thought was in the box before it was opened, they tend to say that they thought from the start that it contained pencils, which suggests that younger children are just as unable to attribute false beliefs to themselves as to others (Gopnik & Astington, 1988).

The false belief-task has also been applied to populations with developmental disorders and different cultural backgrounds. Baron-Cohen, Leslie and Frith (1985) found that the majority of children with high-functioning autism failed in a standard false-belief test, whereas children of equivalent mental age who had Down's syndrome passed the test successfully. As for cultural differences, in 1998, psychologist Angeline Lillard published a comparative review based on ethnographic studies to argue that there are remarkable cultural differences in the practice of folk psychology (see Lillard, 1998). A meta-analysis of false-belief experiments based on 178 separate studies and 591 conditions (Wellman, Cross & Watson, 2001) revealed that the country of children's origin indeed influences performance so that at any one age, children from different countries perform better or worse than each other (e.g. children of a certain age in Australia and Canada perform better than children in the USA and those in Austria and Japan perform somewhat worse). Nonetheless, children in all countries demonstrated roughly the same developmental sequence, which seems to suggest that the developmental trajectory of false-belief understanding is universal across cultures. The meta-analysis (ibid.) also showed that various methodological efforts to make the verbal task easier for children (e.g. by emphasizing that one of the protagonists had a deceptive motive, or by emphasizing the temporal aspects of the situation with the question "When Maxi

comes back, where will he look *first* for his chocolate?") had little effect on the performance.

Although three-year-olds and younger children fail in classical false belief tasks, more recent studies with non-verbal paradigms have produced puzzling results that are currently at the centre of debates on implicit mindreading. Studies which employed the active-helping paradigm indicate that 18-month-olds take another person's false belief into account when tracking the person's action goals (see e.g. Buttelmann, Carpenter, & Tomasello, 2009, Knudsen & Liszkowski, 2012). Experiments with looking-time paradigms even suggest that infants younger than sixteen months may have an implicit understanding of false beliefs (see e.g. Onishi & Baillargeon, 2005, Surian, Caldi, & Sperber, 2007). There is currently little consensus on how to interpret these results (see e.g. the special issue of *British Journal of Developmental Psychology*, edited by Low & Perner, 2012). One attractive option to explain the early success in non-verbal false-belief tasks is to suppose that humans have two cognitive systems for tracking mental states: an early (both developmentally and computationally), fast, and automatic system, and a later developing, more flexible, and more cognitively demanding system (see Apperly & Butterfill, 2009, Apperly, 2011).

The intriguing results of hundreds of false belief studies have fuelled intensive theoretical discussions to which I will turn next. Before introducing the mainstream views on the basic cognitive mechanisms for attributing mental states, a few terminological remarks are in order. Since Premack and Woodruff's seminal paper psychologists have studied the human ability to attribute mental states often under the label "theory of mind." This label is somewhat misleading, however, because it seems to assume that the human ability to attribute mental states is theoretical in nature, when in fact not all researchers share this assumption. A further problem is that the use of "theory of mind" is ambiguous between at least three different meanings. Depending on the context, it is used either as: a) a label for a *particular theory* – theory-theory – that explains the ability to attribute mental states as a theory-driven process (in this sense "theory of mind" refers to a particular *explanans* of the ability to attribute mental states); b) a shorthand for the ability to attribute mental states to oneself and others (in this sense "theory of mind" refers to the main *explanandum* that theories of "theory of mind" seek to explain); or c) an umbrella term for a whole *research paradigm* that takes the ability to attribute mental states as key to human social cognition (in this sense "theory of mind" is used to refer to all versions of theory-theory and simulation theory, see e.g. Leudar & Costall, 2009a or Froese & Gallagher, 2012). In addition, not only does "theory of mind" have several meanings, various other terms are used in parallel to it to designate the ability to attribute mental states. For example, the terms "mind-reading" and "mentalizing" are also used to refer to the ability to attribute mental states. Philosophers also talk about "folk psychology" (sometimes also called "common sense psychology", "naïve psychology", or "everyday psychology"): a conceptual framework for thinking about mental states and processes that ordinary people encounter in their daily lives and talk about in vernacular

terms, such as "pain", "desire", "hope", and "fear." According to Ravenscroft (2010), "folk psychology" is used by philosophers in at least three different senses. It can refer to: a) a set of platitudes about mental phenomena that people typically endorse; b) the quotidian human ability to predict and explain behaviour; or to c) a sub-personally represented mentalistic "theory" of behaviour. The term is common in philosophical discussions on the nature and workings of the mind, especially in relation to *eliminative materialism*, which is the view that our ordinary understanding of the mind as consisting of mental states is a defective theory (Churchland, 1981, 1984). It is less frequently used in the context of empirical research on social cognition.

I prefer to use the term "mindreading" instead of "theory of mind" to refer to the *ability* to attribute mental states. In philosophical literature, mindreading is typically defined as an ability to *explain and predict* the behaviour of agents in terms of their mental states but my definition of "mindreading" is more minimal: it does not require that the attribution of mental states constitute an ability to explain and predict behaviour. I want the definition of mindreading to be open to the possibility that attribution of mental states may have functions besides explaining and predicting behaviour. I will use the term "folk psychology" as little as possible because it is usually taken to be bound to the ordinary use of mental state terms whereas I want to be open to the possibility that people implicitly attribute mental states that have no counterparts in our vocabulary. Concerning "theory of mind", in the articles that constitute the main part of my thesis (appendices 1–3), I have followed Gallagher and Zahavi (2008) by adopting the third sense of "theory of mind": I use it there primarily as a label for the research paradigm that focuses on mindreading, i.e. I use it as an umbrella term to refer to both theory-theory and simulation theory. I later realized that this is not the most fortunate use of the expression. This is because theory-theory and simulation theory are theories about people's *understanding* of the mind rather than theories about the mind. So in the summary article, I prefer to use the term "theories of mindreading" to refer to the generic framework of theory-theory and simulation theory. In any case, I hope that my ways of using the expressions "theory of mind" and "mindreading" are sufficiently clear and easy to follow within the context of this summary.

2.2. The mainstream theories of mindreading

A major aim of research on social cognition is to identify the basic cognitive mechanisms that underlie social understanding and interaction. For the past 30 years, *mindreading* has been considered to be the key to human social cognition. Various theories have been worked out in an effort to explain the underpinnings of mindreading. Most of the theories are versions of either *theory-theory* (TT) or *simulation theory* (ST), or hybrids that combine elements of both. TT, as the label suggests, assumes that our ability to attribute mental states

to ourselves and to others is *theoretical* in nature. According to ST, we use our own psychological machinery to *simulate* the mental states of others.

Social cognition can be studied at different levels of description and explanation. In the mindreading literature, a distinction between personal-level and sub-personal level processes[9] is often made. Simply put, personal level descriptions and explanations pick out features of whole organisms and their behaviour and typically involve the use of intentional vocabulary (e.g., "John believes that p"). Sub-personal level accounts pick out (typically with non-intentional vocabulary, though some authors also permit the use of intentional vocabulary at the sub-personal level) cognitive and neural processes that *underlie* personal level phenomena (e.g., "mirror neurons are activated"). Sub-personal level accounts refer to either functional operations that need to be postulated in order to explain personal-level processes or to neurobiological processes in which the functions are realized. Table 1 lists the main questions that drive research on mindreading at both personal and sub-personal levels of description and explanation.

Table 1.

Level	Personal level	Sub-personal level	
		Functional level	Neurobiological level
Main question	How do individuals attribute and reason about mental states? How is mindreading experienced from the first person perspective? [10]	Which information-processing operations and functional systems need to be postulated to explain mindreading?	What are the neurobiological systems and processes in which these cognitive functions are realized?

Most theories of mindreading provide either personal-level or sub-personal level accounts of mindreading. However, a comprehensive account of mindreading (and of social cognition in general) should be able to cover both levels and also explain how the levels interconnect. A plausible theory of mindreading must minimally: 1) explain the basic mechanisms for both self- and other-directed mental-state ascriptions; 2) unravel the typical development of mindreading abilities; and 3) account for the typical errors people make in everyday mindreading as well as for the pathological breakdowns in mindreading

[9] This distinction was first introduced by Daniel Dennett in *Content and Consciousness* (1969).

[10] "How is mindreading experienced from the first person perspective?" is a personal-level phenomenological question, which is not usually investigated by the standard theories of mindreading. However, any theory of social cognition should be phenomenologically plausible, so I have included the question for completeness.

abilities, such as are manifested in autism and other brain disorders. Any theory of mindreading must obviously be compatible with real-world and experimental data, with scientific knowledge of the neurobiological architecture of the human brain, and with first-person phenomenology of how one experiences instances of mindreading. Furthermore, theories of mindreading need to be evaluated in the context of the wider issue of human social cognition: How much of human social cognition can these theories explain?

In the following, I introduce the two most influential approaches to mindreading: theory-theory and simulation theory.

2.3. Theory-theory

The term "theory-theory" was first introduced by philosopher Adam Morton (1980) (compare Wellman, 1990, p. 129, Stich & Nichols, 1998, p. 422) to label the view that our everyday knowledge of human psychology has the features of a theory. Morton sketched TT in order to contrast it to his own favoured account – "scheme theory." Later on, unlike scheme theory, TT gained many supporters and was elaborated in different ways by various authors.

The main idea of TT is that our everyday attributions of mental states to ourselves and to other people are mediated by a largely implicit *theory* of the functioning of the human mind. To say that our folk psychological competence is theory-like is to say that our mental state concepts are individuated by their functional roles within a framework of inter-related folk psychological principles. A basic explanatory principle could be, for example, a practical syllogism of the form "If a psychological agent wants event y and believes that action x will cause event y, he will do x" (Gopnik & Melzoff, 1997, p. 126). Based on such principles, it is arguably possible to predict and explain the behaviour of individuals.

The roots of TT go back to mid 20th century analytic philosophy. Philosophy of mind witnessed the rise of functionalism in the 1960s – the view that mental states are functional states with causal roles in explanations of behaviour. Wilfrid Sellars (1997, originally published in 1956) outlined the idea that mental states are posits of a commonsense psychological theory or *folk psychology*. David Lewis (1972) developed the idea further and came up with a view that the concepts of mental states are theoretical notions that can be defined by three types of psychological law which differ in what they are about. Laws of the first type are about the connections between observable input and mental states. Laws of the second type are about the interconnections between different types of mental states. Laws of the third type are about the connections between mental states and observable output. These functionalist ideas form the philosophical origins of TT. (See Goldman, 2006, pp. 7–10, Ratcliffe, 2007, pp. 42–46).

Several contemporary analytic philosophers of mind – for example Fodor (1987), Churchland (1981, 1984), Dennett (1987), and Carruthers (1996) –

consider people's everyday reasoning about the mind to be based on a folk-psychological theory about mental states and their interconnections. Whereas Churchland (1981, 1984) argues that our folk psychological theory is fundamentally erroneous and needs to be eliminated and replaced by a neuroscientific theory, other philosophers are less sceptical about its prospects. Either way, in discussions of eliminativism, it is assumed that people in their everyday lives actually use a folk-psychological theory to explain and predict their own and other people's behaviour. Philosophical discussions on folk psychology usually focus their attention on questions about the metaphysics of mind and the meaning of mental state terms. I will not discuss here whether folk psychology as defined by these philosophers is a correct theory (metaphysically speaking), whether it needs to be eliminated[11], or whether mental state terms get their meaning from such a theory. My interest lies in the empirical question of *how human social cognition actually works*, regardless of whether the way people typically understand the nature of the mind is correct or not. Therefore, in the following, I will introduce those versions of TT that have been developed to answer this question. In particular, I will focus upon TT as outlined by cognitive and developmental psychologists such as Alison Gopnik, Alan Leslie, Andrew Meltzoff, Josef Perner, Henry M. Wellmann, and others.

According to TT, the development of mindreading skills is the acquisition of progressively more complex theoretical knowledge (Stueber, 2006, p.106). Two main ways to explain the acquisition of theoretical knowledge have been proposed: either it is a gradual change of theoretical understanding driven by the acquisition of experiential evidence and counter-evidence, or it consists of the maturation of an innate cognitive module (which itself may consist of various sub-modules). Depending on how the cognitive architecture and the development of the folk psychological theory are described, TT has two main versions: the *child-scientist theory* and the *modular theory*.

2.3.1. The Child-Scientist Theory

The philosophical idea that our everyday understanding of the mind is structured like a theory was picked up in the beginning of 1980-ies by developmental psychologists Alison Gopnik, Andrew Meltzoff, Henry M. Wellmann, Josef Perner, John H. Flavell, and others. Since then, countless articles and books have been dedicated to the hypothesis that the processes of cognitive development in children are similar to or even identical with the processes of cognitive development in scientists. The book "Words, Thoughts, and Theories" by

[11] The issue of eliminativism does not seem relevant until it has been proven that people actually use a folk psychological theory to understand others. For example, Ratcliffe (2007, p. 228) argues: "Churchland's eliminativism recommends the elimination of something that was not part of social life to begin with. It is the elimination of what certain philosophers and cognitive scientists think that the 'folk' think, rather than what the 'folk' actually think." Similar arguments have been made by proponents of simulation theory.

Gopnik and Meltzoff (1997) is one of the most elaborate, and probably the most radical, account of the TT hypothesis that has come to be known as the *child-scientist theory*[12]. Gopnik and Meltzoff (1997, p. 32) are careful to point out that the aim of their theory is:

> … not to show that children do science. Instead, we want to argue that the cognitive processes that underlie science are similar to, or indeed identical with, the cognitive processes that underlie much of cognitive development. It is not that children are little scientists but that scientists are big children. Scientific progress is possible because scientists employ cognitive processes that are first seen in very young children.

In other words, Gopnik and Meltzoff don't take the cognitive processes that underlie scientific theory-building to be a late cultural invention. They instead take them to be a product of human evolution. To understand and to evaluate the claim that children's understanding of many aspects of the world is theoretical, we need to know what a theory is. There is a lack of consensus in science and in philosophy of science on what are the basic characteristic features of a theory but Gopnik and Meltzoff (1997, p. 33) alleviate this problem by adopting a conception of a theory that is "as mainstream and middle-of-the road" as possible.[13] They emphasize three kinds of feature that typically characterise theories: structural, functional, and dynamic (ibid. pp. 33–41).

The structural features concern the static properties of theories: theories are *abstract* and *coherent*, they appeal to *causality*, and they make *ontological commitments*. To say that theories are abstract is to say that the vocabulary of the theoretical constructs differs from the vocabulary of the evidence on which the theory is based. In other words, theories involve entities and laws that enable us to explain data instead of simply restating it. Theories are coherent in the sense that entities postulated within a theory are firmly interrelated. The superficial regularities in the data are explained in terms of some causal structure that accounts for the regularities – this is what is meant by the appeal to causality. Finally, theories entail ontological commitments by postulating what exists and how and by generating counterfactual claims. (ibid. pp. 34–36)

When we ask what theories do, we arrive at the *functional features* of theories. According to Gopnik and Meltzoff, theories have three important functions: they enable us to *predict*, to *interpret*, and to *explain* phenomena. Although, in order to predict, we sometimes rely on mere empirical generalizations, theoretical predictions are more specific. Theories allow predictions about evidence that had no role in the initial building of the theory: "A few theoretical entities and laws can lead to a wide variety of unexpected predictions" (ibid. p. 37). Theories also lead to interpretations of evidence,

---

[12] Gopnik (1996) also uses the label "theory-formation theory."

[13] Stich and Nichols (1992, p. 46) use a much looser definition and consider "just about any internally stored body of information about a given domain as an internally represented theory of that domain."

which means that in light of a given theory, some pieces of evidence are more important than others in the context of a particular problem. With regard to what concerns the function of explanation, Gopnik and Meltzoff famously state that "explanation is to cognition as orgasm (at least male orgasm) is to reproduction" (ibid. p. 38, see also Gopnik, 1998, and Gopnik, Meltzoff & Kuhl, 2001, pp. 162–164). In saying this, the authors mean to point out that both a child's and a scientist's search for theoretical knowledge is motivated by the "pleasure" of having an explanation (Gopnik & Meltzoff, 1997, pp. 36–38)[14].

A further important characteristic of theories is that they are *dynamic*: theories change over time under the pressure of new counter-evidence. According to the child-scientist view, theory change in science, as well as in human development, typically involves three phases. At first there is a denial of counter-evidence: the data that speaks against the theory remains overlooked or is considered to be irrelevant. When pieces of counter-evidence accumulate to the extent that they can no longer be ignored, *ad hoc* auxiliary hypotheses are created to explain away the counter-evidence without changing the core of the old theory. Over time, the *ad hoc* hypotheses begin to deteriorate the coherence of the theory, so finally the old theory gets replaced by a new theory. (Gopnik & Meltzoff, 1997, pp. 39–41).

According to the child-scientist hypothesis, all the above mentioned features of theories also apply to children's cognitive structures in various domains: in children's naïve understanding of psychology but also in their early knowledge of biology and physics (ibid. p. 3). Since my focus here is social cognition, I will take a closer look at the child-scientist account of the development of social cognitive abilities.

Gopnik and Meltzoff (1997, pp. 128–134) argue that social cognition relies on theoretical knowledge from the start: already newborn infants possess something like a "starting state theory" (see also Gopnik & Wellman, 1992, p. 169) about themselves and other people that is represented in an abstract and coherent way. Simply put, they claim that infants come to the world with an implicit theory that they are like other people in some important respects. On the basis of infant imitation studies (Meltzoff & Moore, 1977, 1983, and 1989), Gopnik and Meltzoff (1997, p. 129) write:

> … from birth, information about action that comes, literally, from inside ourselves is coded in the same way as information that comes from observing the behavior of others. There is a fundamental cross-modal representational system that connects self and other.

In other words, from birth, information acquired about one's own body is generalized to others and vice versa (ibid. p. 132). The innate mapping between the visually perceived motions of others and the infant's own kinaesthetic sensations enables smooth affective and temporal coordination between the infant

[14] Perner (1991, p. 244) provides a less metaphorical account of explanation: to explain is to provide a mechanism (or model) that underlies the observed causal sequence of events.

and its caregivers. Obviously, the psychophysical states that infants are sensitive to are not as sophisticated as the mental states at the centre of the adult theory of mind: "Rather than having a concept of psychological agents, young infants seem to have a concept of persons, which combines mind and body" (ibid. p. 133). What makes this early form of cognition theory-like? Gopnik and Meltzoff argue that theoretical structures can be inferred from the fact that infants respond distinctively when people's behaviour contradicts their implicit theoretical predictions (ibid.). Unfortunately the authors leave unspecified how they come to the conclusion that infants make any theoretical predictions in the first place. It remains doubtful that infants' early understanding of action carries the necessary structural and functional features of a theory – the newborn's ability to imitate certain facial gestures seems to involve sensory-motor and affective abilities rather than any cognitive or "theoretical" features. Thus it seems unwarranted to attribute to newborn infants the ability to make theoretical predictions just because they react distinctively to different kinds of input. Yet this does not seem to worry Gopnik and Meltzoff.

According to Gopnik and Meltzoff, infants distinguish people and objects from birth but it takes time before the ability to distinguish what acts are effective on people from what acts are effective on objects develops. Infants begin to make sense of the basic features of physical and psychological causality by around nine months. This change in infant understanding is considered to be a theoretical shift that leads to an understanding that "the first set of entities [people] is susceptible to communicative acts while the second [objects] is subject to actions that involve spatial contact" (ibid. p. 145). An ability to understand the role of communicative interaction is displayed when infants begin to initiate communicative intentional behaviours such as pointing. Communicative gestures indicate that children no longer conceive of actions as mere bodily movements but instead begin to acknowledge that actions are intentional – i.e. they are directed towards objects or other people (ibid. pp. 138–145).

By around eighteen months infants become interested in regularities that characterize unsuccessful attempts to influence other people's behaviour; especially cases of conflict between one's own goals and those of another person. Infants seem to particularly enjoy doing things that the caregivers have prohibited them from doing: they curiously explore the boundaries of what they are allowed to do. By this time, infants seem to understand that other people can have desires that contradict their desires. (ibid. p. 149) This new competence is demonstrated for example in the nonverbal "cracker-broccoli" test (Repacholi & Gopnik, 1997): when another person has previously expressed a desire for broccoli, 18-month-old infants gave her broccoli although infants themselves typically preferred crackers. In contrast, 14-month-olds solved the task egocentrically, giving the other person food they themselves preferred. This ability is linked to the understanding of the goal-directedness of behaviour: 18-month-olds can infer the particular goals of failed attempts, thus differentiating intentions from outer behaviour. (Gopnik & Meltzoff, 1997, p. 150).

Between the second and the third year of life, infants arguably develop a mentalistic, albeit not yet representational theory of *desire*. By this time, children have learned that people tend to strive for what they desire, that they are disappointed when their desires are unfulfilled and happy when they are fulfilled, and that there may be conflicts between their own desires and those of other people (ibid. pp. 157–158). However, during this period, children still fail to understand that mental states can *represent* the world differently from how things really are, as demonstrated by their inability to pass false belief tasks. Before children start to understand the representational aspects of the mind, they treat perception and desire as simple causal links between the mind and the world (Gopnik & Wellmann, 1992, p. 150).

Around the age of three, children understand also *beliefs* as direct links to states of affairs in the world. "This view has variously been called a 'copy theory' (Wellman, 1990), a 'Gibsonian theory' (Astington & Gopnik, 1991), a 'situation theory' (Perner, 1991), or a 'cognitive connection' (Flavell, 1988) theory of belief" (Gopnik & Wellmann, 1992, p. 151). This early theory of belief leads children to specific behavioural errors, as when they insist that they thought from the very beginning that a Smarties box was filled with pencils (which are shown in fact to be there), even though they had at the outset expressed a belief that it contained candies (which the appearance of the container initially suggested) (Gopnik & Astington, 1988).

According to Gopnik and her colleagues (Gopnik & Meltzoff, 1997, Gopnik & Wellman, 1992), between three and five years of age, children's understanding of the mind changes from a non-representational theory to a representational theory. Arguably, this change has all the typical features of a theory change. At first, children tend to ignore counter-evidence to the early simple causal theory of the connection between the mind and the world. Later they create auxiliary hypotheses to account for the phenomena that contradict the earlier theory but preserve the basic structure of the old theory; for example, they occasionally recognize that beliefs may misrepresent but ignore the influence of false beliefs on action. Finally, usually by the time of their fifth birthday, children have replaced the simple causal theory with a representational theory of mental states. (See Gopnik & Wellmann, 1992, pp. 151–153).

The account outlined by Gopnik and Meltzoff (1997) is one of the best known versions of the child-scientist theory. Other advocates of the child-scientist approach have elaborated similar views but there are differences in many of the details of the different versions of TT. For example, Perner (1991, p. 251) does not think that the causal theory of mental states (which he calls both "(mentalistic) theory of behaviour" and "situational theory") will finally be *replaced* by a representational theory of mind: "There is no simple replacement since … we stay situation theorists at heart." Perner (ibid. p. 252) argues that around the age of four, the situational theory is *extended* rather than supplanted by the representational theory:

> Even as adults we remain situation theorists whenever possible and treat mental states as straight propositional attitudes. However, in contrast to young children, we are also able to take a representational view when necessary – to explain cases of misrepresentation, for instance.

This means that whereas Gopnik considers theory change in children to be analogous to the replacement of the Ptolemaic astronomical theory by Kepler's theory, Perner compares it with the extension of classical genetics into molecular genetics (ibid. pp. 251–252).

For Henry Wellman (1990, p. 127) our everyday theory of mind constitutes a *framework theory* (or a paradigm in Kuhn's sense) rather than a *specific theory*. Framework theories constrain the formulation of specific theories by establishing the domain of explanation and by defining what counts as relevant evidence (ibid. p. 136). According to Wellman (1990), children make an ontological distinction between mental entities and physical entities from a very early age but they acquire a first theory of mind at around three years of age. This claim derives from a conceptual constraint: Wellman considers the ability to attribute representational states to be necessary for having a theory of mind. However, this does not mean that children lack any understanding of mental phenomena before the age of three. Two-year olds arguably have what Wellman calls a *simple desire psychology*: they understand desires not as relations to propositions, but as relations to actual objects or events in the world. Such an understanding of desires is intentional (desires are directed at something) but not representational (see Wellman, 1990, pp. 210–212).

According to Wellman, having a theory of mind begins with a transition from a *simple desire psychology* to a *belief-desire psychology* that in his account takes place roughly around the age of three years. In contrast, according to Perner (1988), Flavell (1988) and Gopnik & Meltzoff (1997), a representational understanding of mental states appears about a year or a year-and-a-half later than Wellman claims (see Wellman, 1990, p. 243). The difference here between the views of Wellman and others lies in whether the earliest so called "direct copy theory" of beliefs is considered to be representational or not. Wellman argues that three-year olds have an initial understanding of representations because they demonstrate an understanding of the difference between fictional and reality-oriented mental states (i.e. between imaginings and beliefs) (ibid. p. 255). However, initially they have a "hit-or-miss type of understanding of misrepresentation," which explains why they fail in false belief tests (ibid. p. 254). Three-year olds understand the mind as containing direct copies of objects in the world; between the age of three and six, they progressively develop an interpretive understanding of representations (ibid.). This process is part of a larger shift from a passive understanding of the mind to an active understanding of the mind: "the earliest theory of mind – that of three-year-olds – can be seen as having a containerlike nature; the later theory of mind … can be seen as having a homunculuslike nature" (ibid. pp. 268–269).

In short, Wellman emphasizes two theoretical shifts in the development of theory of mind in early childhood: the shift from simple desire psychology to an initial belief-desire psychology (Wellmann labels it "a copy-container theory of mind"), and the shift from an initial belief-desire psychology to a later belief-desire psychology (also labelled "the interpretive-homuncular theory" by Wellman) (ibid. p. 278).

In sum, although there are differences in the details of the versions of the child-scientist theory, there is a substantial amount of agreement among the proponents of this approach. The advocates of the child-scientist approach agree that children's understanding of the mind has a theory-like structure. The basic theory of mind is not innate but develops gradually through acquired experience and is equally applicable to self and other. There are several steps of theory-change in the preschool years and the most important change involves a shift from a non-representational conception of mental states to a representational conception.

Obviously, the development of theory of mind does not end by the age of six but empirical and experimental evidence that support the child-scientist theory usually come from the studies of early childhood. Much less empirical research has been carried out on mindreading in adolescence and in adulthood, which is why the debates on the mature theory of mind have taken place rarely and more on purely philosophical grounds. As Henry Wellman (1990, p. 97) writes: "it is surprising that so little effort has been spent on describing what the [adult's] theory is, beyond asserting that it must exist." He is one of the few psychologists to have sketched what an adult's theory of mind might be like but he is careful to add that his sketch "has not emerged from a program of empirical research" (ibid.).

It is often pointed out that because no-one has been able to explicitly formulate more than a few illustrative examples of folk-psychological principles, it is doubtful that any such principles are actually represented by the human cognitive system. Of course, nobody expects that in order to apply a theory of mind, people need to be able to spell out the rules of it explicitly. There is an analogy between mindreading and linguistic skills: children become competent language users long before they know anything about the rules of grammar, presumably because linguistic rules are implicitly represented in their cognitive system. The difference is, however, that the rules of grammar for many languages have been explicitly formulated by linguists as a result of a systematic study of language but this is not the case for mindreading. It is unclear whether the principles upon which mindreading rests can be explicitly formulated. Some proponents of TT have suggested that the theory underlying mindreading is not represented in the form of laws or principles but rather in the form of *theoretical models* (Maibom, 2003, 2007, Godfrey-Smith, 2005; see also Newen & Schlicht, 2009). If the model-based version of TT is on the right

track, it would explain why it is so difficult to spell out any rules of theory of mind: theory of mind simply is *not* represented in the form of rules.[15]

Another issue raised by the critics is the question of *self-directed mind-reading*. According to TT, self-directed mental state attributions are arrived at by the same cognitive processes that underlie mental state attributions to others: the same generic folk psychological theory is applied in both cases. This seems counter-intuitive, however. It appears that we have a more direct acquaintance with the contents of our own mind than with the mental states of other people. It does not seem very plausible that in order to know what mental states I currently have, I need to make inferences that are similar to those that I have to make in the case of third-person mindreading. A theory-theorist reply is that although usually we do not make *explicit* inferences when we attribute mental states to ourselves, *implicit* inferences nevertheless play an important role in this process (Gopnik, 1993, Carruthers, 1996, p. 36). The impression of knowing directly or non-inferentially one's own mental states can be explained by drawing an analogy with the "illusion of expertise" (Gopnik, 1993, pp. 10–12). A chess master may no longer need to explicitly consider the positions of the pieces on the chess board to estimate the situation: the phenomenology of an expert may be simply one of seeing that, for example, the queen is vulnerable. Since every mindreader is an expert on reading one's own mind, we may also have the experience of directly accessing our own mental states, although it is actually a result of swift and implicit inferences. This reply is not particularly convincing, though. Theory may have a role in conceptualizing the mental states that we experience from the first person perspective, and TT may explain how one reasons about one's past or future mental states, but it is unlikely that one's current mental states have to be inferred in the same manner as other people's mental states are inferred according to TT: from perceptual information. Until theory-theorists come up with a more specific account of detecting one's current mental states, I have to agree with Nichols and Stich (2003, p. 158), who state that it is "simply preposterous to suggest that the reports people make about their own mental states are being inferred from perceptions of their own behaviour and information stored in memory." An account of self-directed mindreading must be able to explain how a person can report her current thoughts and other mental states even when there is no perceivable behavioural evidence of them. TT provides no such explanation (ibid).

Another serious problem for TT is related to the so-called *frame problem*. The frame problem, which first arose in the context of artificial intelligence, is the issue of how a cognitive system is able to compute what information is relevant for a particular task in any given context (see Shanahan, 2009). In the context of TT, the problem is to explain how the human cognitive system is able to swiftly pick up pieces of relevant information for determining the pertinent

[15] If mindreading is explained as a process where we use our own mind as a model to attribute mental states then we have arrived at another theory of mindreading: simulation theory.

mental states to explain and predict individual behaviour in any given context, without having to compute vast amounts of information available to the system (see e.g. Heal, 1996). This problem is closely related to the issue of *holism*: behaviour is never a result of a single mental state but instead arises from complex combinations of mental states.[16] For example, the fact that I take an umbrella with me when I go out can be explained by my belief that it may start raining but only if I also have the desire to remain dry when it rains and if I believe that an umbrella protects me from getting wet. In principle, any behavioural act might be driven by almost any mental state given some combination of other mental states, which makes deciding what mental states to attribute in order to predict or explain a particular act incredibly difficult, if not impossible. Yet, in most everyday contexts, people mindread with considerable ease.

TT faces several other challenges as well but, because of scope limits, I brought out only the most important ones. It is, however, worth mentioning that the child-scientist theory bares the burden of explaining how children growing up in very different social, cultural, and physical environments, come up with the same theory of mind around the same age. This is an issue that has inspired theory-theorists to produce a different version of TT to which I will turn next: the modular approach.

2.3.2. The Modularist Theory

According to the child-scientist theory, children acquire a theory of mind in the same vein adults acquire scientific theories: by gathering evidence and searching for the best way to explain data. But whereas most theories in various fields (e.g. astronomy) are not universally shared, the basic theory of mind seems to be roughly identical across all normally developing humans (see Segal, 1996, pp. 152–153, Scholl & Leslie, 1999, p. 137). How do children across the world come up with the same fundamental theory of mind, around the same age? The answer of the child-scientist theory is largely empiricist, with a small nativist component. It is argued that: 1) children come to the world with the same basic knowledge – newborns possess a preliminary theoretical understanding of other people as entities "like me"; 2) children all over the world are exposed to similar patterns of evidence and counter-evidence concerning other people's behaviour, and they are thereby pushed towards reinventing the same theory of mind (Gopnik, 1996, pp. 172–174). The critics of the child-scientist theory, however, doubt that the developmental path to acquiring the same theory of mind can be explained away so easily:

> … it is surely not the case that if one collected a few million scientists who started out with the same initial theory, then gave them the same counter-evidence, that nearly all of them would arrive at the same revised theory – within roughly the same time span" (Segal, 1996, p. 153).

[16] This line of thought applies if one buys into some version of realism of mental states.

An alternative way to explain the universal developmental trajectory of theory of mind is to shift the explanation more towards the nativist side at the expense of the empiricist component. This move is made by modularist theory-theorists who explain the development of the ability to mindread as a result of the maturation of an innate cognitive mechanism or module, consisting of several sub-modules. Before outlining the most prominent modular accounts of mindreading, let me briefly introduce the concept of modularity in general.

The concept of modularity was outlined in detail by Jerry Fodor (1983) and it soon became an important topic in philosophy of mind, cognitive psychology, and evolutionary psychology. According to Fodor (1983), modular systems have all or most of the following features: domain specificity – modules have the function of solving problems in particular domains; mandatory operation – when an input is received, it is automatically processed in a particular way independent of voluntary control; limited central access to the mental representations that input systems compute – the flow of information *out of* a mechanism is restricted; high speed of computations; informational encapsulation – the flow of information *into* a mechanism is restricted; 'shallowness' of outputs; fixed neural architecture; specific breakdown patterns; particular temporal pace and sequence in ontogeny. Although Fodor (1983) originally applied the idea of modularity only to low-level systems, such as those that underlie perceptual abilities and language processing, and argued that higher-level cognition that leads to belief-formation consists of general-purpose processes, others have extended the idea of modularity to high-level cognition as well. For example, Carruthers (2006) has argued that the human mind is massively modular but he has considerably relaxed the definition of modularity by giving up some of the necessary features of the classical Fodorian modules, such as informational encapsulation (Robbins, 2010). As a result, modularity has become a rather slippery concept and in any given context, one needs to specify which definition of modularity one has in mind: a system may be considered modular in one but not in some other sense.

Segal (1996) distinguishes between *synchronic* and *diachronic* dimensions of modularity: whereas synchronic modularity characterizes modules as static devices, diachronic modularity characterizes modules as they develop over time following genetically determined trajectories.[17] Segal brings out four different notions of synchronic modularity: *intentional modularity*, *computational modularity*, *Fodorian modularity*, and *neural modularity* (ibid.). A similar distinction is made between the following three conceptions of modularity: *the epistemic conception*, *the algorithmic conception*, and *the hardware conception* (Samuels, 1998, Gerrans, 2002).

[17] E.g. Chomsky's account of the development of the language faculty can be interpreted as involving diachronic modularity. Language learning in childhood can be explained as a process of *parametrization* (setting of parameters) within genetically predetermined limits: according to the input received from the language environment, children implicitly adopt the rules of that particular language. (Segal, 1996, p. 146)

Segal's notion of *intentional modularity* refers to a mental mechanism that is described in purely intentional vocabulary and is characterized by either informational encapsulation, limited accessibility, or both (see Segal, 1996, pp. 142–143). It bares similarities to the notion of *epistemic modularity*, which is defined as a "domain-specific body of innate knowledge" (Gerrans, 2002, p. 307)[18].

Computational modularity characterises a computational system that turns input representations into output representations via the use of a set of syntactic rules (see Segal, 1996, pp. 143–145). This notion is similar to the *algorithmic conception* of modularity (see Samuels, 1998, p. 580, Gerrans, 2002, p. 307), according to which modules are defined by their computational properties. Whereas every computational module plausibly realizes an intentional module, the reverse does not hold (Segal, 1996, p. 144): an intentional module or a domain-specific body of knowledge may exist in a system that only contains domain-general computational mechanisms (see Samuels, 1998, p. 583).

A *Fodorian module* is, according to Segal, a special case of the computational module, since it involves certain additional properties. In other words: while every Fodorian module is a computational module, not every computational module is Fodorian. I have already outlined the nine typical features of a Fodorian module above (but see also Segal, 1996, p. 145).

Finally, Segal's *neural module* (similar to the *hardware conception* of Samuels (1998, p. 579) and Gerrans (2002, p. 307)) is a functional part of the brain that can be described in purely neurological vocabulary. Any of the other three modules may be realized in neural modules, but they need not be, since intentional, computational, or Fodorian modules may also be realized by distributed features of the brain. At the same time, neural modules do not necessarily realize intentional or computational modules because a capacity may be fully explained by referring to neural processes only, without implying that the neural processes satisfy any computational or intentional descriptions.

Let us now look at how the concept of modularity has been used in the context of social cognition, particularly in developmental psychology. Like most developmental cognitive psychologists, the defenders of the modular account of theory of mind are interested in those human information processing systems that form the basis for cognitive development. In contrast to the child-scientist theory, which construes the development of mindreading as a domain-general process of theory-building, modularists argue that natural selection has produced specific mechanisms that drive the development of theory of mind. The modularity account of the theory of mind was first proposed in order to explain the double-dissociation between the presence of an ability to mindread and the level of domain-general intelligence: Baron-Cohen, Leslie, and Frith (1985) discovered that in the case of high-functioning childhood autism, the

[18] Whereas Samuels (1998) considers domain-specific bodies of innate knowledge to be possibly non-modular knowledge-structures, Gerrans (2002) classifies such structures as modular.

ability to mindread (as measured with the standard false-belief task) was selectively impaired, whereas in the case of severely retarded children with Down's syndrome, it was selectively intact.[19] In the following, I introduce two of the most prominent versions of the modular TT accounts, Leslie's and Baron-Cohen's, and then consider some most prominent arguments for and against the modular view.

Alan Leslie defends a computationally (algorithmically) modular account of mindreading (see Scholl & Leslie, 1999, p. 133, Gerrans, 2002, p. 308). Instead of picturing the child as a great scientist who passes through radical theory shifts under the pressure of empirical evidence, Leslie (1994) characterises the ontogenetic development of theory of mind as the maturing of a domain-specific system. He distinguishes between three sub-mechanisms within this system. These are: *ToBy* (Theory of Body Mechanism) for processing mechanical agency; *ToMM1* (system 1 of the Theory of Mind Mechanism) for processing actional agency; and *ToMM2* (system 2 of the Theory of Mind Mechanism) for processing attitudinal agency.

The first mechanism of Leslie's account, ToBy, arguably begins to develop at approximately 3 or 4 months of age (Leslie, 1994, p. 140). Its purpose is to track three-dimensional moving objects in the environment, to distinguish between two categories of such objects (agents and non-agent objects), and to compute their mechanical properties. The module receives input from two distinct systems of visual processing: from the system that underlies recognition of three-dimensional objects and from the system that analyses motion. ToBy attributes mechanical properties to objects by employing a primitive notion that Leslie labels FORCE. When an object begins to move on its own (i.e. its source of FORCE is within the object), it is automatically categorized as "agent"; when it begins to move as a result of physical contact with another moving object (it receives FORCE from an external source), it is categorized as "non-agent object." For example, when a child witnesses a moving object colliding with another object and the launching of the second object immediately after the collision, ToBy automatically assigns them complementary mechanical roles. (See Leslie, 1994, pp. 123–137)

Whereas Leslie introduces ToBy as a module for tracking mechanical features of bodies, he postulates a distinct mechanism – ToMM – as the seat of the child's mindreading abilities. ToMM computes intentional features of agents and consists of two subsystems: ToMM1 is concerned with goal-directed actions and ToMM2 processes propositional attitudes and meta-representations. (See Leslie, 1987, p. 140)

[19] Similar results were found in a follow-up study (Baron-Cohen, Leslie & Frith, 1985) where high-functioning autistic children, children with Down's syndrome, and normally developed children were tested for their mechanical, behavioural, and intentional understanding of picture stories: autistic children performed much worse than the control groups in tasks involving intentional understanding but they were successful in tasks involving mechanical and behavioral understanding.

ToMM1 underlies the ability to understand that, unlike non-agent objects, agents *perceive* their environment and *aim for goals*. Whereas non-agent objects can only be moved via direct physical contact, agents can also be set to move by factors that are at distance in time and/or space. Around 6 to 8 months, infants begin to attend to the direction of the gaze of their caregivers. By the end of the first year, they become competent gaze followers, being able to accurately track what the other person is looking at. At the same period of time they begin to display instrumental "requesting" and "refusing" behaviours, thereby communicating their own volitional states to adults. Leslie argues that these behaviours indicate the maturation of the ToMM1 module, which enables children to grasp simple intentional relations between agents and distant objects or states of affairs. (See Leslie, 1994, pp. 140–141)

Finally, between 18 and 24 months of age, ToMM2 comes online (ibid. pp. 141–142). ToMM2 underlies the ability to create *metarepresentations* and to understand agents' intentional relations to fictional states of affairs (ibid.). This module is first put to use in children's production and understanding of *pretend play* (Leslie, 1987). Pretence involves two simultaneous representations of the same situation. By pretending, for example, that my furry winter hat is a cat, I am not confused about what kind of object it literally is: I know that it is a hat but I treat it as if it was a pet – for example, by stroking it and by giving it imaginary milk from an imaginary bowl. Thus when one understands or produces pretence, one must simultaneously represent what the object or the situation literally is and what it is pretended to be. Leslie (1987) argues that children exhibit an ability to understand that the same state of affairs can be represented in different ways as early as they start to pretend – typically by their second birthday. If Leslie's account is correct, a question arises: why are children unable to solve the false belief task until the age of four, although they are able to metarepresent other's mental states as early as by the age of two? Leslie (1987, pp. 422–423) argues that attribution of false beliefs is considerably more complex than attribution of secondary mental states in pretence. In the case of pretence, what the situation is pretended to be is made obvious by the pretender (e.g. one holds a banana close to one's ear and speaks to it as if it were a telephone), whereas in the false belief task, what the person falsely believes about the situation needs to be worked out by the observer (Leslie, 1987). According to Leslie (ibid.), children younger than four-years old fail the false belief test because they are unable to select the correct content for the false belief. Leslie and colleagues (see e.g. Leslie & Roth, 1993, Scholl & Leslie, 1999) argue that because beliefs tend to be true, the ToMM automatically imposes the current situation as a default content for any ascribable belief. In the case of false belief, however, the default interpretation needs to be overcome. For this job, Leslie postulates yet another special mechanism: *SP* (Selection Processing). SP is a presumably a non-modular cognitive device that enables the mindreader to choose between various possible contents by taking relevant background information into account; it inhibits, when necessary, the default contents that the ToMM automatically

produces (Scholl & Leslie 1999, 147). The upshot is that a full-fledged mindreading competence may require both domain-specific and domain-general processes. This conclusion is fully compatible with Leslie's modular account because he holds the view that "it is not that the entirety of ToM is modular, but only that ToM has a specific innate basis" (Scholl & Leslie, 1999, p. 134).

Simon Baron-Cohen (1997) is the author of another well-known modular account of mindreading. He proposes that at least four mechanisms – *IT* (Intentionality Detector), *EDD* (Eye Direction Detector), *SAM* (Shared Attention Mechanism), and *ToMM* (Theory of Mind Mechanism) – underlie the human capacity to mindread. Baron-Cohen's account of mindreading is modular in the strongest possible sense: it unites the epistemic, the architectural, and the hardware conceptions of modularity (Gerrans, 2002, p. 308).

The first mechanism ID (Intentionality Detector) is an amodal perceptual module which has the purpose of attributing goal-directedness to selected stimuli. It gets activated whenever one perceives something as an agent – most typically when one perceives an object with self-propelled motion. ID is similar to Leslie's ToBy but whereas Leslie described ToBy as having the function to compute mechanical properties in general, Baron-Cohen's ID processes only properties of agents (Baron-Cohen, 1997, pp. 32–38).

The next mechanism EDD (Eye-Direction Detector) is a module within the human visual system dedicated to processing information about what other agents are seeing. It automatically registers the presence of eyes (or eye-like stimuli), enables one to follow the direction of the gaze of another and to infer what the agent is looking at. In normally developing individuals, perceived eye-contact (i.e. you and another agent are looking at each-other) triggers physiological arousal which is typically experienced as pleasant (ibid. pp. 38–44).

ID and EDD enable infants to track other people's goals and perceptual states by creating *dyadic representations* in the form of "agent desires x" or "agent sees x." In order for the child to be able to know that the same object is the focus of attention of both oneself and someone else, one needs to compare one's own perceptual content with another individual's perceptual content. This comparison is achieved via a third mechanism: SAM (Shared-Attention Mechanism). SAM comes online towards the end of the first year of life. It usually receives inputs from EDD (although it can also use information from other modalities) and transforms them into *triadic representations*; i.e. it represents relations between oneself, another agent, and an object (which may be another person). SAM also serves as a link between EDD and ID and thereby enables the interpretation of another's eye direction in terms of volitional mental states (ibid. pp. 44–50).

Finally, Baron-Cohen postulates a fourth mechanism – ToMM (Theory of Mind Mechanism). He borrows the name of the module from Leslie and declares an agreement "with much of what Leslie says about the workings of ToMM" (ibid. p. 51). Whereas the previous three modules were limited to processing only simple volitional and perceptual states, ToMM enables

attribution of the full range of mental states. It represents epistemic (or propositional) mental states such as pretending, knowing and believing, and develops over time into a cognitive mechanism that provides a coherent implicit theory of how different mental states relate to each other and to actions. Baron-Cohen argues that in order to mature, ToMM needs to receive input from SAM – its development begins by converting the triadic representations received from SAM into metarepresentations. Since the functioning of SAM implies the functioning of ID and EDD, the development of ToMM presupposes that the three other modules already function properly (ibid. pp. 51–55).

Modular accounts of mindreading differ in many respects – from what kind of concept of modularity is assumed to how many sub-mechanisms are postulated and how their functioning is explained – but they are all opposed to the idea underlying the child-scientist theory that the development of children's mindreading capacities depends only on the domain-general ability of theory-building. Instead, mindreading is assumed to rely on the maturation of innate and domain-specific cognitive mechanisms. This has led some of the critics to accuse the modular accounts of "anti-developmentalism" (Gopnik, 1996, p. 174, Gopnik & Meltzoff, 1997, p. 54). For example, Gopnik and Wellman (1992, p. 284) argue that modular theories are unable to account for the typical sequence of conceptual changes that characterize the development of theory of mind, such as the replacement of an early non-representational understanding of mental states by a later representational account. They claim that unlike standard modular accounts, the child-scientist theory provides a dynamic picture of development and acknowledges the importance of the role of the interactions between the child and its environment: theories change in a systematic order under the pressure of new evidence.

Scholl and Leslie (1999) ward off the critique by emphasising that nothing requires modules to be developmentally static. There are restrictions to the information flow into and/or outside of a module but this does not mean that modules cannot develop internally. One way to explain how modules develop is to adopt Segal's (1996) account of *parametrization* of diachronic modules. Children in different language environments end up speaking different languages. This can be explained as a result of a diachronic process of adjusting the parameters according to the current language environment within the genetically specified limits of the language module. However, Scholl and Leslie (1999) prefer the view that the mindreading module is not subject to parametrization, because the end result of the maturation of the theory of mind module is, in their view, a uniform theory across all normally developing humans. They explain the early changes in theory of mind as a result of additional modules coming on-line and changes in later development as performance improvements, rather than changes in competence (see Scholl & Leslie, 1999).

If both the child-scientist theory and the modular view can equally well explain the development of mindreading, how can we decide which of the two approaches is correct? A general argument in favour of the modular approach

appeals to the speed of everyday social cognition. Most everyday social situations require an ability to make swift judgements about other's mental states and the modular account enables us to explain how such fast processing is computationally possible. Modularity enables us to avoid the *frame problem* (see section 2.3.1.) by preventing the system from searching the whole cognitive space for information that is relevant for mindreading in any particular case. This is one of the reasons why modular processing is faster than domain-general processing (Gerrans, 2002, p. 310). A more specific argument for the modular view rests on its ability to explain why children with high-functioning autism with normal IQ (i.e. with intact domain-general cognitive abilities) fail the false-belief test (Baron-Cohen, Leslie & Frith, 1985). The child-scientist theory seems incapable of accounting for the developmental dissociation between mindreading and domain-general intelligence. However, if mindreading develops as a module independently from general intelligence, the dissociation seems to find a rather straightforward explanation.

Things are more complicated than that, however. Currie and Sterelny (2000, p. 149) argue that mindreading is unlikely to be strongly modular (in the sense of being the product of only informationally encapsulated cognitive mechanisms) because, in order to formulate beliefs about other people's mental states, specific contextual information needs to be (and often is) taken into account. However, they defend what they call a modest modular account of mindreading: the idea is that there is a modularized layer of information processing between perception and social belief fixation, where socially relevant "tags" are added to the perceptual inputs (ibid. p. 154). For example, when we see a good actor in tears, playing the role of a heartbroken character, we automatically and involuntarily perceive the scene as of seeing a sad person but we do not thereby necessarily come to *believe* that the actor is actually sad, since the background information ("this is a play") blocks this inference. If the process of social belief fixation itself was modular in the strong sense of being informationally encapsulated, we would not be able to override the perceptual impression and would automatically end up with a belief that the person *is* sad. The modest modular account seems to adopt the advantages of both the child-scientist theory and the strongly modular theory, while at the same time avoiding their weaknesses: it escapes the general frame problem by introducing a modular level of information processing, and avoids the excessive rigidity that is typical of fully modular systems.

There are additional reasons to favour a modestly modular explanation that regards mindreading proper as a domain-general capacity that takes as inputs the results of early-processing modules. When the early processing modules fail to do their job, the domain-general system does not receive the necessary input to successfully compute mental states, which explains why high-functioning autistic patients have great difficulties with mindreading. The lack of an ability to mindread is, however, not the sole characteristic of autism – autism is a syndrome that is associated with a number of symptoms, including sensory-motor problems and abnormalities in perceptual processing (Gerrans, 2002, p.

315). A modestly modular account of mindreading may offer a more parsimonious explanation of a wider range of autistic symptoms, including problems with mindreading, whereas the hypothesis that autism is the result of a damaged mindreading module requires additional explanations for other autistic symptoms, such as problems with sensory processing.[20]

A modest version of Baron-Cohen's theory would mean that ID, EDD, and SAM may indeed be independent modules, whereas mindreading proper is a result of domain-general theorising. In fact, although Gerrans (2002, p. 308) claims that Baron-Cohen's account of mindreading is overall modular in a strong sense, this may not hold for all four modules of the theory: Baron-Cohen indeed argues that ID, EDD, and SAM are likely candidates of classical innate modules but he concedes that the fourth mechanism, ToMM, may be more open to learning (Baron-Cohen, 1997, p. 57). Leslie's account is equally flexible in this respect: Leslie allows the mechanism for selecting the relevant contents for mental-state ascriptions to be domain-general.

In sum, the issue of whether mindreading is a result of modular or domain-general capacities is not a black-or-white question. It is plausible that mindreading is a result of both, domain-specific and domain-general processes. The two branches of TT – the child-scientist approach and the modular approach – need not be understood as mutually exclusive. Instead elements of the child-scientist theory and the modest modular approach can be combined to overcome some of the problems that each approach faces alone.

2.4. Simulation theory

> *You know my methods in such cases, Watson. I put myself in the man's place, and, having first gauged his intelligence, I try to imagine how I should myself have proceeded under the same circumstances.*
>
> Arthur Conan Doyle, 1894[21]

For a few years after the publication of Premack and Woodruff's 1978 paper, the discussion over mindreading mainly consisted of a dispute between different versions of TT. In 1986, a radically different approach to mindreading emerged, based on the claim that instead of relying on theoretical knowledge about the

[20] The modest modularity thesis seems to leave unexplained the success in false belief tasks of children with Down's syndrome, whose domain-general reasoning is impaired, (Baron-Cohen, Leslie & Frith, 1985, 1986). However, other studies have shown that children with Down's syndrome actually have difficulties with mindreading, which suggests that passing a false-belief test does not ensure that one is able to successfully mindread in real life contexts, and that mindreading is not completely independent from domain-general reasoning abilities (see e.g. Yirmiya et al., 1996).

[21] Quotation via Gordon (1986).

laws of the functioning of the human mind, people use their own minds as *models for simulating* or *replicating* the mental processes of other people in order to make sense of their behaviour. The basic assumption of ST is that in order to attribute mental states to others, people rely at least partly on the same mental processes that underlie their own psychological states. Let's assume that John wants to simulate a decision making process of Mary. According to ST, the process proceeds roughly as follows: 1) imaginary premises (those that Mary presumably has) are fed into John's decision making system; 2) the system forms a decision based on the given information; 3) the outcome is projected onto Mary in the form of a mental state attribution. But is it not the case that John needs to rely on theoretical knowledge in order to decide which premises need to be fed into the simulation mechanism to begin with? If this is so, ST seems to collapse into a version of TT. Defenders of pure simulationism reject the view that theoretical knowledge (in the sense assumed by TT) is required for mental simulation. In what follows, I will give an overview of various versions of ST to outline how simulationists have tackled various problems that a theory of mindreading is expected to solve. I will focus especially on Alvin Goldman's account because it is the most elaborated version of simulation-based theory of mindreading to date.

2.4.1. The early versions of simulation theory: Heal and Gordon

Let us first look at the earliest publications where ST was for the first time sketched as an alternative to TT: articles by Jane Heal (1986), and Robert Gordon (1986). In "Replication and functionalism," Heal (1986) contrasted the *functional strategy* that she considered to be at the heart of TT with what she labelled a *replicative strategy*. She argued that the functionalist TT is unlikely to underlie everyday social understanding because of the problem of holism: in a functional framework, any action could in principle result from many different combinations of mental states. To decide what mental states are relevant to explain a certain behaviour in a particular case requires extremely complex theorising (if it can be done at all), yet humans are able to mindread with considerable ease and speed. This makes it highly implausible that in order to mindread, people rely on the functional strategy, especially if there is a much easier explanation of how mindreading may be accomplished. The easier explanation, according to Heal, goes roughly as follows: people imagine the world from another person's point of view, they have their cognitive system process the input from such imagination, they then attribute the result to the person whose behaviour they wish to predict or explain. In order to do this, they need no complex folk psychological theory – all they need is to have a working mind and an assumption that other minds function in a similar manner to theirs. Heal argued that to get to the initial state that serves as an input for the replication process one does not need to make any functionalist-style theoretical inferences. She argued that, instead, the input states can be created by looking at or thinking about the *world* that surrounds the other and by assuming that others *qua*

thinkers are like oneself, rather than by theorising about the insides of other people's minds (Heal, 1986, p. 137). She did not exclude the possibility that some pieces of theoretical knowledge are adopted in this initial stage – for example, one may need to use one's knowledge of optical laws to figure out what is perceivable from the other's perspective – but she emphasized that such knowledge is different from the use of folk-psychological laws about mental states and their interactions.

In the same year, Robert Gordon (1986) published the paper "Folk psychology as simulation" where he outlined similar ideas. He started the paper by arguing that the best way to *predict* what I will do is to *decide* what I will do: in such cases we rely on *hypothetico-practical reasoning* instead of making nomological inferences. He went on to argue that one can also predict what other people are about to do by *simulating* their decision making process. For example, to *predict* what move my chess opponent will make next, I take up my opponent's perspective, decide what move to make, and attribute that decision to the opponent. A similar process takes place when one engages in *retrodicting* and *explaining* the actions that have already occurred: I imagine having acted like the other person and consider what mental states this behaviour may have resulted from.

Obviously, this process needs to include adjustments for the relevant differences between myself and the other person – otherwise the result would characterize myself instead of the person of interest. Sometimes it is practically impossible to make such adjustments accurately, for example when an amateur chess player tries to simulate what is going on in the mind of a professional player (compare Goldman, 1989, p. 83). At other times it is rather easy, especially if the other person is very similar to me in relevant respects. According to Gordon, the ability to carry out necessary adjustments for mental simulation develops in early childhood and increases with age. He explains the fact that three-year-olds typically fail the standard false belief test by arguing that they simulate egocentrically, whereas older children pass the test presumably because they have learned to make relevant adjustments by taking the other's perspective into account (Gordon, 1986, p. 168). Gordon also pointed out that ST provides a more viable explanation than TT of why high-functioning autistic children have difficulties with the false-belief task whereas children with Down's syndrome are able to pass it: it is not because the former lack and the latter possess an ability to theorise about mental phenomena but because mindreading is based on a specific (and possibly modular) ability to carry out mental simulations (ibid. p. 169). To prevent a potential phenomenological argument against ST that simulation is unlikely to be the fundamental means of mindreading because we rarely ever experience carrying out a simulation routine, he adopts the view that simulations usually run implicitly, below the level of awareness (ibid. p. 170).

2.4.2. Varieties of simulation

After the appearance of the early articles on mental simulation introduced above, several other researchers also picked up the idea that mindreading may be based on processes of simulation rather than theorising. Different accounts of ST were outlined, for example, in a double special issue of *Mind & Language* in 1992, which was dedicated to the debate between ST and TT. The general underlying idea that mindreading is *process-driven* (i.e. the same processes that underlie our own mental states also underlie attributions of mental states to others) rather than *theory-driven*, was cashed out in various ways. As Nichols and Stich (2003, p. 132) note, in the early days of ST, the discussion on mind-reading was framed as a battle between two camps – TT and ST. This created a situation in which proponents of ST tried to tag the label "simulation" to just about any process that might be understood as simulation in some sense of the word, despite the fact that many of these processes had little or nothing in common. The concept of mental simulation was put to use to capture processes that take place on a personal level (e.g. involving voluntary imagination and projection from oneself to another, see Goldman, 1989), as well as for processes that take place on a sub-personal level (involving automatic activation of mirror neurons and other neural processes, see Gallese & Goldman, 1998). Goldman (2006) distinguishes between low-level and high-level mindreading which apply to different types of mental states and are explained by different models of the simulation process, also allowing theoretical elements to play a role in mindreading. Some authors assume that simulation involves introspection and projection from oneself to another person (Goldman, 1992, 2006). Others explicitly deny that introspection plays a role in simulation and appeal to "imaginative identification" instead of projection from oneself to the other person (Gordon, 1995). Although most authors consider ST to be an empirical hypothesis about the cognitive processes underlying mindreading, Jane Heal (1995, 1998) took the route of defending the idea of simulation as an *a priori* thesis concerning the recreation of the contents of mental states of others in mindreading. Her main point was that in order to think about another person's thinking (in the broad sense of covering all propositional attitudes), one needs to think directly about the subject matter of that person's thinking.[22] Going into the details of all of the different versions of ST[23] would exceed the scope of this chapter. For this reason I'm going to focus on the most prominent version of ST – that of Alvin Goldman. Whereas earlier versions of ST were directly opposed to TT, more recently elements of TT and ST have been combined to create hybrid theories (Heal, 1995, Nichols & Stich, 2003, Goldman, 2006), and this is also the strategy that Goldman has recently used.

[22] Heal (1998, p. 483) coined the term "co-cognition" for "thinking about the same subject matter."

[23] Stich and Nichols (1997, p. 299) have even argued that the diversity among the different versions of ST is so vast that the term "simulation" has become futile: "It picks out no natural or theoretically interesting category."

2.4.3. Alvin Goldman's account: From pure ST to a ST-TT hybrid

No overview of ST can ignore the contribution of Alvin Goldman. As Frédérique de Vignemont (2009, p. 457) has written about him: "No philosopher has done more to display the resourcefulness of mental simulation." Goldman has developed and transformed ideas on mental simulation for over two decades, arriving at a detailed ST-TT hybrid (with an emphasis on simulation) with his 2006 book "Simulating Minds".

His early ideas (Goldman, 1989) were similar to those of Heal (1986) and Gordon (1986): he considered simulation to be an implicit or explicit process of mentally putting oneself into the other's situation, generating further states from the pretend input, and projecting the result to the target. In his earlier writings, he attempted to avoid bringing in elements of TT, arguing that in order to create input for simulation, one has to rely on non-theoretical knowledge of the other person's perceptual situation and may assume that the other has the same "basic likings or cravings" as oneself (Goldman, 1989, p. 82). *Contra* Dennett's (1987) claim that any simulation necessarily involves theoretical knowledge, Goldman argued that mental simulation can work without theorising if the interpreter's input states are relevantly similar to those of the interpretee, and if the processes driving the simulation are isomorphic to the processes producing mental states in the interpretee (Goldman, 1989, p. 85). According to Goldman, this is indeed the case for human mindreaders: in the course of mental simulation, relevantly similar pretend-states are created in our cognitive system and processed by the same mechanisms that process genuine mental states of the interpretee (ibid. pp. 85–86). In other words, while naturally produced mental states like ordinary beliefs and desires give rise to other ordinary mental states (and often also to behaviours), pretended or imagined mental states lead to mentalistic predictions. Whereas according to Leslie (1994), the ability to pretend is an early manifestation of a theory of mind, Goldman argues that it may instead reveal an early ability to simulate (ibid. p. 87). But how do simulators manage to take into account the differences between themselves and the people they mindread? Goldman concedes that people need to use some empirical information to accommodate interpersonal differences but denies that such information needs to be mediated by a folk-psychological theory. He adds that his aim is not to claim that mental state attributions are always accurate: people often make errors while mindreading, and typical errors reflect an egocentric bias, which is another reason to support ST over TT (see also Goldman, 2006, p. 148).

The beginning of the 1990s witnessed the discovery of *mirror neurons* in macaque monkeys: these are neurons that are activated both when an individual executes an action and when one observes a similar action in others (Di Pellegrino et al., 1992, Rizzolatti et al., 1996). Thereafter, indirect and direct evidence has been found for the existence of mirror neurons in humans (see Fadiga et al., 1995, Iacoboni & Mazziotta, 2007, Mukamel et al., 2010). Goldman was quick to establish theoretical links between this neuroscientific discovery and the simulationist approach to social cognition. With Vittorio Gallese, one of the discoverers of the mirror neurons, he argued that the

function of mirror neurons might be mental simulation (Gallese & Goldman, 1998). Because macaque monkeys have very limited if any mindreading abilities, they carefully suggested that mirror neurons "represent a primitive version, or possibly a precursor in phylogeny, of a simulation heuristic that might underlie mind-reading." (ibid. pp. 497–498)

Later on, Goldman distinguished between *low-level* and *high-level mind-reading*[24] in his hybrid theory. Therein he laid emphasis on simulation processes (Goldman, 2006). He describes low-level mindreading as a "primitive" form of mindreading in the sense of being cognitively relatively simple (it presumably relies on mirroring processes and enables us to recognize types of simple mental states without identifying propositional contents) and in the sense of having evolved earlier than high-level mindreading (Goldman, 2006, p. 113). According to Goldman, low-level mindreading underlies face- and body-based emotion recognition as well as the attribution of feelings and intentions. He reviews empirical evidence showing that subjects who have difficulties with *experiencing* a certain type of emotion also have problems *recognizing* the same type of emotion in others, whereas their ability to experience and to recognize other types of emotions is preserved. For example, patients with damaged amygdalae, who have lost the ability to experience fear, are also impaired in their ability to recognize fear in others (ibid. pp. 115–116). Similar paired results are found in patients with damaged insula and basal ganglia who no longer experience nor recognize disgust, and in subjects whose experience of anger is temporarily blocked by a drug called sulpiride. Goldman argues that it is difficult to see how TT could explain such selective deficits, because it is unlikely that one can have an otherwise intact theory of emotions minus knowledge of just one type of emotion; ST, on the contrary, not only incorporates these selective deficits, but even predicts them. To explain how emotion recognition proceeds in detail, Goldman considers several different computational models for simulation-based low-level mindreading (see ibid. pp. 124–132). He favours a model of "unmediated resonance" according to which when one perceives another person's facial expressions, ones own neural substrate of the corresponding emotion is triggered, classified, and attributed to the target (ibid. pp. 127–129). As this model demonstrates, mirroring alone is not sufficient for mindreading: the result of the mirroring also needs to be classified and imputed to the target. Moreover, the mental state that occurs as a result of mirroring must match the state of the target, but at the same time it has to function differently from it, because otherwise it would result in emotional contagion rather than recognition of another's mental state. Moving from emotions to actions, Goldman argues that in the case of both genuine and mirrored intentions to act, the brain creates specific motor plans. The difference between actual and mirrored motor plans is that in the case of mirroring, the motor plans are inhibited from triggering real actions. They translate into

[24] For the criticism of the distinction between low-level and high-level mindreading, see De Vignemont, 2009.

predictions of actions instead. If this model accurately describes action mirroring, we can speculate that *echopraxia* (involuntary repetition of another person's actions) may result from an impaired ability to inhibit the mirrored action plans (see also Goldman's discussion of a patient who cannot but produce the movements he imagines, ibid. p. 160).

The critics have pointed out that the use of the concept of simulation to characterize the activity of mirror neurons and other neural resonance processes is unjustified. The main critique of low-level ST is not provided by advocates of TT, but by authors critical of both, ST and TT (see chapter 3). For example, Gallagher and Zahavi (2008, pp. 178–180, see also Gallagher, 2008c) argue that the function of mirror neurons should be understood in terms of *enactive perception* in preparation for social interaction instead of *simulation* (see also Froese & Gallagher, 2012, p. 448). They draw on two common definitions of simulation and argue that what mirror neurons do satisfies neither of the definitions. According to *pretence definition*, to simulate means to pretend. But who is pretending in the case of mirror neurons? It cannot be the person, because the person has no awareness of, not to mention conscious control over, the activity of mirror neurons. Gallagher and Zahavi argue that it also makes no sense to say that the brain is pretending, because the brain is not an agent: "As vehicles, neurons either fire or don't fire. They don't pretend to fire." (Gallagher & Zahavi, 2008, p. 180) According to the *instrumental definition*, simulation is understood as an activity of using something as a model for something else. Once more, Gallagher and Zahavi insist, the person cannot *use* mirror neurons in any way, and it is absurd to use this concept for sub-personal processes. Herschbach (2008), *au contraire*, supports the simulation interpretation, arguing that "brain mechanisms could 'use' other brain mechanisms as models without requiring the intelligence of a person."[25] In any case, it is important to acknowledge that the simulation interpretation is not the only possible interpretation of the workings of mirror neurons, although it is the most prevalent one. But let me now return to Goldman's account.

Whereas in Goldman's account, low-level mindreading targets emotions, feelings and intentions and is "comparatively simple, primitive, automatic, and largely below the level of consciousness" (Goldman, 2006, p. 113), high-level mindreading targets more complex mental states such as propositional attitudes, and it may be to some degree conscious and partly under voluntary control (ibid. p. 147). However, Goldman is also quick to concede that processes that underlie high-level mindreading may most of the time be completely implicit (ibid. p. 151). In high-level mindreading, *pretend-states* play a central role: the mindreader creates in herself pretend-states, feeds them into her cognitive system, and attributes the results to the target. What are pretend-states and how are they produced? According to Goldman, pretend-states are mental states of the interpreter that have the purpose or function of replicating or matching in

[25] Slors (2010) prefers the concept of "neural resonance" instead of speaking about simulation or mere perception.

relevant respects the mental states of the target of mindreading (ibid. p. 149, see also p. 37). These states are created by *enactive imagination* (E-imagination) which is a process that enables the endogeneous (and sometimes voluntary) production of mental states that are otherwise produced exogeneously (ibid. p. 149). In contrast to the mere supposing (or S-imagination) that characterizes detached hypothetical reasoning, E-imagination affects the subject's mind more deeply and engages largely the same psychological processes as would be operative if one were really in the imagined circumstances (ibid. p. 175). For example, although any act of seeing requires a visual stimulus from the environment, one can create a similar quasi-visual experience by retrieving the necessary input from memory. Accurate E-imagination requires not only a general capacity for E-imagination, but also task-specific knowledge that can be used to build up pretend-states (ibid.). For example, if I have no idea what an elephant looks like, I cannot accurately E-imagine an elephant. To E-imagine something, there must be resemblance between a naturally occurring mental state and a corresponding pretend-state. The resemblance may, but does not have to, appear at a phenomenological level; for Goldman it suffices to have the resemblance at a functional or at a neural level (ibid. p. 158).

What reasons are there to believe that relevant similarities actually exist between ordinary mental states and E-imagined states? Goldman puts forward several pieces of evidence. For example, patients with unilateral visual neglect (as a result of brain damage) experience a "gap" in a certain part of their visual field that manifests both while seeing and while imagining a scene, which suggests that *seeing* and *visualizing* rely on (partly) overlapping neural processes (ibid. p. 152–157). There is analogous evidence for shared neural underpinnings in motor imagery and in motor production (ibid. p. 157–160). In addition, Goldman claims that besides visual and motor imagery, E-imagination extends to purely conceptual domains where imagistic or motor properties are lacking (ibid. p. 160–162). Such broad use of the concept of E-imagination serves the purpose of enabling us to explain how pretend propositional attitudes (pretend-beliefs and pretend-desires) can be created as initial input for the simulation process.

Unfortunately, the nature of *conceptual E-imagination* remains rather obscure and the empirical findings Goldman uses to support his thesis of conceptual E-imagination are unconvincing. He draws on studies that demonstrate that *the bystander apathy effect*[26] occurs not only when other people are present, but also when one merely imagines other people being present (ibid. p. 161). But it remains unclear how this example speaks for the existence of *conceptual E-imagination* rather than a mere supposition conjoined with elements of visual E-imagination: imagining other people is surely not a case of pure conceptual thinking but seems to involve quasi-visual (or quasi-perceptual) features. The example of a famous priming effect seems to fare no

[26] The bystander apathy effect is the tendency to refrain from helping a person if other people who could also help are present.

better. When people are primed with words that are associated with the stereotype of elderly people, they are inclined to walk more slowly (Bargh, Chen & Burrows, 1996). Goldman argues that this might be a case of conceptual E-imagination: people who have been primed with the elderly stereotype may imagine being elderly themselves and slow down their pace as a result. However, there is no direct evidence that people thus primed imagine being elderly themselves. Furthermore, the priming effect itself has recently been called into doubt: Doyen et al. (2012) were not able to reproduce the original results of the Bargh et al. (1996) study. An even more serious fault of Goldman than citing unconvincing evidence is that he does not sufficiently clarify the concept of conceptual E-imagination. He mentions that the similarities between real and pretend propositional attitudes are to be found at the functional level (Goldman, 2006, p. 161), but leaves the reader wondering about what exactly this means: he does not elaborate on how pretend-desires and pretend-beliefs match with and at the same time differ from naturally produced desires and beliefs. One option to back up Goldman's theory would be to claim that pretend-states and naturally produced states share a similar content, but differ in properties of "non-content" (see Heal, 1996). In this case, a pretend-state is not some "faint copy" of the genuine state but, instead, a particular way of representing its content (ibid.). In order to know what content is a suitable input for simulation, some theoretical information must probably be involved, but that would be no problem for a hybrid account.

Tooming (2013) has recently argued that pretend desires cannot be distinguished from naturally produced desires and that therefore they cannot form a separate kind of mental state. He proposes an alternative model of high-level mindreading which operates without pretend-states: mindreaders use their own genuine mental states as input for the simulation process and rely on theoretical information to adjust the *result of* the simulation process to match the mental states of the target (ibid.). This is an interesting idea but Goldman would probably reject it because it puts too much weight on theorising. In Goldman's hybrid theory, theorising has a secondary role as compared with the service performed by simulation: in cases of predictive mindreading, theoretical information is used to adjust the input of simulation so that it matches the other person's initial mental states; in cases of retrodictive mindreading, theorising creates hypotheses that are "tested" via simulation (Goldman, 2006, pp.183–185). Simulation based on the genuine mental states of the mindreader may work in cases where the current mental states of the simulator are fairly similar to the relevant mental states of the target, but it is hard to see how it could work as an effective strategy in all or most cases of mindreading. For example, if I want to predict what someone who does not share my fear of heights is going to do on top of the Eiffel tower, it makes little sense to run a simulation based on my fearful state first, and to subsequently adjust the result of the simulation. It would be more economic to make relevant inferences without first running the simulation (by reasoning, for example, that the subject is likely to enjoy the panorama and to take pictures of Paris, because that is what people without

acrophobia usually do on top of the Eiffel tower). Alternatively, I may try to imagine what it's like *not* to be afraid of heights (i.e. I create a pretend state), feed this imagination to my system, and impute the result to the target. In the first case, theorising is doing all the work; in the second case, pretend states (in the sense of mental states that go beyond the repertoire of my mental states) are involved. One might argue that this example is no serious objection to Tooming's model because only rarely will the mindreader be incapable of having a mental state that characterizes the target. This is not the point, however. What matters is that the *current* mental states of the mindreader and the target often differ. Recreating a mental state (even by retrieving one's own experience from memory, for example) *for the purpose of mindreading* is a special case of mental state production. Goldman could argue that if a mental state has been produced for the purpose of matching the state in the target, we are already dealing with a pretend-state rather than a genuine mental state of the mindreader. So it seems that if a simulationist account of high-level mind-reading is at all feasible, Goldman's model is preferable to the "pretenceless" model sketched by Tooming (2013), given that a more detailed and convincing account of pretend states is provided, such as that suggested by Heal (1996). Alternatively, high-level mindreading may rely more heavily on processes that do not involve simulation, such as folk-psychological theorising.

Besides pretend states, another important aspect of Goldman's theory is its reliance on *introspection*. Although introspection is not appropriate as a scientific method, it seems obvious that people have some sort of introspective access to their current mental states: for example, if I don't express a random thought in any way, it is virtually impossible for others to attribute it to me but I can easily attribute it to myself (see Goldman, 2006, p. 230). Goldman argues that introspection as a specific mechanism for detecting one's current mental states is a necessary element of mindreading: in order to attribute a simulated mental state to the target, the mental state must first be recognized and classified by the mindreader.

Not all simulation theorists agree with Goldman on the role of introspection. Gordon (1995) argues that introspection or inference from oneself to another person is unnecessary for mindreading. In Gordon's model, simulation involves a "re-centring" of the egocentric perspective: I transform my perspective to match the perspective of my target. I thereby begin to use the pronoun "I" to refer to the target instead of to refer to myself. As a result, projection of the mental state from myself to the other person becomes redundant because the simulated mental state is linked to the other person from the start. But how does a person know what mental states she currently has, either genuinely or as a simulator? In Gordon's model (see Gordon, 1996, pp. 15–16), there are no specific introspective processes; instead, one detects one's own mental states via an *ascent routine*. In order to answer e.g. the question "Do I believe that Tallinn is the capital of Estonia" one simply asks oneself "Is Tallinn the capital of Estonia?" – if the answer is affirmative, then one attributes the belief to oneself. However, the problem with this idea is that it only works with beliefs but not

with other mental states such as, for instance, hopes or fears: I cannot tell whether I *hope* that it will rain tomorrow by asking "Will it rain?" (see Nichols & Stitch, 2003, p. 194).

A related difference between Goldman's and Gordon's account is the role of mental concepts in their respective theories of mindreading. According to Goldman, mindreading via simulation requires that one be able to appropriately classify mental states, i.e. it presupposes that one already possesses mental state concepts (see Goldman, 2006, ch. 10). Things are the other way around in Gordon's model, where simulation does not require the possession of mental concepts; rather "our ability to grasp the concepts of mind and the various mental states depends on our having the capacity to simulate others" (Gordon, 1996, p. 11).

In Goldman's view (2006, p. 186), Gordon confuses two issues: the issue of who is the subject of the states that result from the process of simulation, and the issue of what are the contents and the tags associated with the contents of those states. Within the content of simulation, the pronoun "I" may indeed refer to the target of the simulation, but since one cannot *literally* transform into another subject, the state still needs to be introspectively identified, classified, and attached to the target. The pretend state is a state *of* the subject, even when the subject attributes it to someone else. As Goldman explains it, there is a "difference between the (pretend) state of deciding to do m, and the final genuine state of believing that the target will decide to do m" (ibid. p. 187). To proceed from the former to the latter, i.e. from a representation to a metarepresentation, one needs to first classify the state via some process of self-monitoring and further attribute it to the target (ibid.).

To support the view that mindreading involves introspection, Goldman refers to neuroscientific evidence that shows there to be increased activity during mindreading in brain areas that are associated with self-reflection (ibid. p. 162–164). If Goldman is right about the role of introspection, it suggests that introspective identification of one's own mental states could occur without an ability to attribute mental states to others, whereas other-directed mindreading cannot function without introspection. In contrast, TT predicts that mental state attribution to self and mental state attribution to others go hand in hand because any mental state attribution draws on the same folk psychological theory (see e.g. Carruthers, 1996, p. 36). Although young children who fail in standard false belief tasks also fail to attribute false beliefs to themselves (Gopnik & Astington, 1988), it does not show that they lack introspective access to their *current* mental states (see also Harris, 1992, p. 132). The attribution of a false belief to oneself is not a case of introspection, but a case of third-person mindreading targeted at one's no longer held mental state. Empirical evidence seems to pull toward Goldman's side: there is data of first-person mindreading in cases of impaired third-person mindreading (Goldman, 2006, p. 224 and pp. 236–237). This suggests that self-directed attributions of current mental states rely on special processes that differ from third-person mindreading (ibid.) As for the explanation of how introspection works, Goldman does not provide a

detailed account but he hypothesises that neural properties underlying the current mental states of the subject are processed, represented, and classified in mental-state terms by some dedicated cognitive processes (ibid. p. 251). He also describes introspection as a perception-like process in that it operates via attention: I do not attribute to myself every mental state that I am having but I attribute those current conscious states to which I attend (ibid. pp. 242–245).

2.4.4. Some general challenges for ST

I have introduced several versions of ST and outlined Alvin Goldman's hybrid theory in somewhat greater detail. Discussing specific counterarguments to different versions of ST would exceed the scope of this introduction. However, I would like to point out some general challenges that apply to most versions of ST. This will lead me to elaborate on why Goldman's hybrid approach is more attractive than pure ST.

Firstly, as already acknowledged by early simulation theorists (see e.g. Heal, 1986, pp. 138–139), before one is in any position to simulate one's own or someone else's mental states, one already has to have some insight into the mental states of the target in order to choose the appropriate input for the simulation process. Most simulation accounts explain how one predicts one's own and other people's possible future mental states by appealing to beliefs about the other's current mental states. They run into trouble in explaining how the simulation process gets started in the first place. So the problem is that it is not possible to simulate from scratch: to simulate, one already has to have some knowledge about the target's mental states. Proponents of ST have tried to overcome this problem in various ways but their solutions are either unconvincing or raise the suspicion that some form of theorising is in play. This is presumably one of the main reasons why Goldman gave up his pure version of ST and turned into a ST-TT hybrid theorist instead: in his model, the input states are chosen with the help of theoretical knowledge. Goldman's recent hybrid seems to be more prominent than any pure simulation account, which also suggests that a pure ST might be unfeasible.

There is no problem of choosing correct input when it comes to low-level simulation because it is assumed that perceiving another person's facial and bodily gestures automatically triggers the necessary neural processes that produce matching states in the observer. So a possible solution to the problem of high-level simulation would be to argue that low-level simulation provides input for high-level simulation. However, low-level mindreading includes categorization of mental states which requires that one possesses mental state concepts, and everyone except for Gordon agrees that mental state concepts cannot rely on pure simulation, so some form of theoretical information is likely to be involved even in the categorization phase of low-level mindreading.

Another challenge for most versions of ST is to explain what ensures that the results of the simulation process characterize the target rather than the simulator. It seems that if we can know other minds only on the basis of our own minds,

we are locked to our own minds and can never reach to someone truly different from us. Thus any simulation account needs to explain how interpersonal differences are taken into account in the simulation process. Here again, it is difficult to come up with an explanation for discrepant mental state attribution without appealing to one's ability to make theoretical inferences about mental states. Not surprisingly, Goldman's solution to this problem includes bringing in elements of TT. The issue is different in regard to low-level simulation, where mental states matching those of the target are directly triggered via specific neural processes (e.g. mirror neurons). But here we have a different issue: the need to explain why we attribute the mirrored mental state to the target rather than experience it as our own. ST assumes that, in both cases, the underlying neural substrate is largely the same. The critics argue that conceptualizing the sub-personal processes underlying low-level mindreading in terms of simulation is problematic. They insist that it makes more sense to conceptualize them in terms of enactive perception (see e.g. Gallagher, 2007).

A further issue is that ST needs to explain how several different, and even contradictory, mental states – especially affective states – can be simultaneously present in a single cognitive system when these states rely on largely overlapping sub-personal processes. Gallagher (2008c) uses an example of seeing that an entomologist is overjoyed to hold an ugly arthropod, when at the same time one is experiencing repugnance towards the creature. He argues that ST is unable to adequately explain how one can accurately capture the mental state of the other person, because it is implied that in order to attribute a mental state to a target, the mental state of the mindreader must match that of the target. In the given example, neither the emotional nor the motor state of the mindreader matches that of the entomologist. A similar problem occurs in social situations which require attributions of distinct mental states to more than one person at the same time, which may be the case in many social situations that involve paying attention to more than one person. For example, when I see two children fighting over a toy, I may have a simultaneous recognition that one of them is angry and the other one is miserable, and I may in addition experience a third emotion – for example, irritation as a reaction to what I see. Assuming that although it seems that one attributes the states simultaneously, one actually rapidly switches between different mental states, is not likely to solve the issue, because emotions require some temporal duration to be experienced as emotions. At the end of the day, it is an empirical issue whether it is possible to simulate a mental state that contradicts the simulator's current mental state, or whether it is possible to simulate several different mental states at the same time. But as far as I know, simulation theorists have not even recognized this as a problem. This is presumably because the majority of empirical support to ST comes from simple lab experiments, where the social input is seriously impoverished compared to real life social situations. To my knowledge, there are no experiments where the experiment subject would simulate more than one person at a time.

2.5. The hybrid account of Nichols and Stich

Goldman's "Simulating Minds" (2006) was not the first hybrid account that integrates theory-like and simulation-like components. Three years before Goldman published his most recent monograph, Shaun Nichols and Stephen Stich outlined what they characterized as an "eclectic"[27] account of mindreading in their book "Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds." They combined selected elements of TT (both in its child-scientist and modularist versions), ST, and some original ideas of their own that have no counterparts in the standard theories of mindreading. Stich and Nichols argued that "eclecticism" is a good thing because it allows us to provide more specific explanations of a wider spectrum of the issues related to mindreading than any of the unified theories that are bound to one central idea.

The approach of Nichols and Stich is "boxological" – they outline a model of the possible architecture of a mindreader's cognitive system by postulating a set of functional units that carry out specific tasks in the system. They focus on explaining the basic cognitive architecture of three tightly related mental phenomena: pretence, mindreading, and self-knowledge. I will here focus on their account of mindreading, which forms the most central part of the book. The term "boxology" comes from the strategy of graphically depicting functional units as boxes in a larger schema that includes a complex system of such special-purpose units and arrows that mark the direction of information flow in the system. The authors hardly say anything about how the various functional units postulated by the theory might be implemented in a biological organism, but this is characteristic also to TT, for example. The book also provides a "just so story" – an account of how the mindreading system may have evolved. The whole account is highly speculative in a good sense of the word, and the speculative nature of it is explicitly embraced by the authors: Nichols and Stich (2003, p. 200) openly declare that they expect that in the light of new evidence and constructive critique, parts of the theory will be proven mistaken. They express a hope that other parts can be improved and further elaborated. Unfortunately, the theory has not received as much critical and constructive feedback as the authors hoped for, so a refined and empirically elaborated version of it has not emerged. However, since it was the first systematic hybrid account of mindreading, it is worth giving a brief overview.

Nichols and Stich (2003, pp. 13–15) start by outlining what they call *the basic architecture of the mind*. They assume that the mind is a system which contains, broadly speaking, two types of representational states: *beliefs* and *desires*.[28] These states differ functionally: they are brought about in different

[27] Nichols and Stich (2003) characterize their account as "eclectic"on pages 11, 18, 59, 60, 100, and 148 of the book.

[28] See Heal (2005, p. 183) for a critique of Nichols and Stich' clear-cut distinction between belief and desires, and Ratcliffe (2007, pp. 187–221) for a general critique of belief-desire psychology.

ways and they interact differently with other components of the mind. Beliefs are caused either by perceptual processes or inferred from already existing beliefs. In the boxological model, the *Belief Box* gets input from the *Perception Box* and from *Inference Mechanisms*, which include a special mechanism called the *Planner* that enables one to figure out ways of achieving one's goals. Desires typically arise from *Body-Monitoring Systems* or other (unspecified) *Desire-Generating Mechanisms* that feed into the *Desire Box*, or they are formed as sub-elements of broader desires. The information typically flows as follows: there is an input from perceptual and inferential mechanisms into beliefs and from desire-generating mechanisms into desires. Beliefs and desires feed into practical reasoning, i.e. they provide input to the *Decision Making System*, which in turn has feedback loops back to beliefs and desires. From the Decision Making System, the information passes on to *Action Control Systems*, which finally generate behavioural output. Based on this model, Nichols and Stich begin to build up a mindreading system by adding various special mechanisms to the basic architecture of the mind.

First they add a special mechanism that enables hypothetical reasoning – the *Possible World Box* (PWB). The PWB enables an individual to make more flexible behavioural choices by predicting the results of different actions without carrying those actions out in reality. It works roughly as follows: one feeds a representation of "I will do x" to the PWB and "clamps" it, adds one's beliefs from the Belief Box to the PWB, and updates the whole system (with the help of the *UpDater*, of course), thus generating a representation of a situation that would have emerged if the organism had actually done x. This representation will serve as a prediction of what would happen if one actually did x. Although designed by evolution for the purpose of making better decisions, the PWB becomes an *exaptation* – its function shifts in the course of evolution so that the PWB also enables one to predict another organism's actions. Nichols and Stich assume that our evolutionary ancestors had a preliminary concept of "goal," which, jointly with the PWB, enabled our ancestors to make simple predictions of what another individual is likely to do. (Nichols & Stich, 2003, pp. 62–64) This is not yet a system for mindreading, however – it merely enables one to predict actions.

The evolution of an early mindreading system is described as follows. Gradually, *Desire Detection Mechanisms* evolve: our ancestors become capable of detecting other individual's desires by using various behavioural cues, facial expressions, and later on also verbal expressions (ibid. pp. 78–80). The mind already contains a *Planner* – a special mechanism for computing the most optimal ways for achieving goals. The Planner becomes another exaptation; it originally evolved to enable the organism to plan its own actions but now gets recruited to represent the action plans of others (ibid. pp. 80–81). Also a third mechanism comes into play – the *Mindreading Coordinator*. The Mindreading Coordinator mediates information between the Belief Box and the Planner: it gathers relevant beliefs about the target's goals and desires and "orders" plausible plans for achieving those goals from the Planner. It subsequently

generates a prediction that the target is likely to follow (one of) the plan(s). The Coordinator also generates beliefs about the sub-goals or instrumental desires of the target. (ibid. pp. 81–83)

Although Nichols and Stich advise against using the term "simulation" (ibid. p. 133) because the term is used to denote many different processes and thereby creates a great deal of confusion, they point out that several elements of their model are borrowed from ST. For example, the idea that the cognitive mechanisms that enable us to make inferences and plan actions are recruited to predict other individuals' inferences and actions is congenial to most versions of ST. Another simulation-like element in the theory is the process they call "default belief attribution." Default belief attribution means that when another person's mental states are modelled in one's PWB, one's own beliefs about the world are copied to the PWB as a default strategy (ibid. p. 85). However, this brings along one of the central problems of ST: one also needs to take into account interpersonal differences, thus the system needs to be able to overcome default attributions in appropriate ways. Solution: Nichols and Stich add *Discrepant Belief Attribution Mechanisms* to the schema (see ibid. pp. 88–92). This is where theory-like elements come into play. The mindreader needs to rely on a rich body of knowledge – a theory – to make inferences about discrepant beliefs. One of the simplest types of discrepant belief attribution is the inference of beliefs from perceptual states: one can infer, for example, that because the other person did not see that an object is in location x, she does not believe that the object is in that location. But this is not always true – the target may have acquired the belief that the object is in location x in some other fashion – for example, someone may have told her. So the mindreader gradually obtains a rich body of knowledge on how different types of mental state are interrelated and linked to behaviour. This idea is inspired by the child-scientist version of TT.

The hybrid model also borrows an element from modularist theories. Namely, the authors hypothesise that mindreading is triggered by some dedicated modular mechanism, such as an *Agency Detector*, which might be a separate mechanism or a part of Desire Detection Mechanisms.

A theory of mindreading must be able to explain also self-directed mindreading. According to Stich and Nichols, one uses the same mindreading system in order to attribute mental states to others as well as to predict one's own possible future mental states. But the authors convincingly argue that the theory-theorist idea that one needs to make inferences based on behavioural data to know one's current thoughts and feelings is absurd, and they postulate a special Monitoring Mechanism that enables one to detect one's current mental states. (See ibid. pp. 150–199)

Besides combining elements of ST and TT, the most important original contribution of Nichols and Stich (2003) is the postulation of the PWB – a specific mental workspace in which rich descriptions of what the world would be like if the assumptions were true are built from a few assumptions. Hypothetically, the PWB first evolved to enable agents to make better decisions via

hypothetical reasoning but it was later recruited also to enable the prediction of actions by other agents. With some additional functional elements, it evolved into a full-blown mindreading system. What seems to be a weakness of the model (see figure 3.4 in Nichols & Stich 2003, p. 87) is that the PWB gets direct input from the mindreader's Belief Box but not from the Desire Box. In this respect, ST (e.g. as in Goldman's version) seems to provide a more parsimonious and intuitively more plausible explanation, since it allows the mindreader's cognitive system to create matching states (pretend-desires) rather than mere beliefs about desires. A related issue is that in Nichols and Stich's model, people attribute to others by default only the content of their Belief Box but not the content of their Desire Box. However, there is evidence that people also tend to default-attribute to others their preferences and other desire-like states (Goldman, 2006, p. 167). Perhaps this issue can be easily fixed by adding a direct link between the PWB and the Desire Box.

Another issue is that a clear-cut distinction between beliefs and desires is problematic – for example it would be difficult to classify *evaluations* as belonging to either category (Heal, 2005). Thus another possible solution might involve merging the Belief Box and the Desire Box. However, this would in turn require serious modifications to other parts of the model (ibid.).

An equally serious worry is that Nichols and Stich do not explain how it is possible to attribute mental states that have no representational content, such as moods or somatic states (e.g. pains or itches). In their model, mindreading can proceed only via the PWB but the PWB is designed to compute mental states only in the format of propositional attitudes. Projecting a direct link from the Body-Monitoring Systems to the PWB is not likely to solve the issue because the PWB cannot handle mental states other than propositional attitudes.

In sum, Stich and Nichols provide a complex boxological model of how the mindreading system may be built from simpler functional units. The model combines elements of TT and ST, as well as some original ideas. Although it demonstrated how a hybrid account has the potential to explain a wider range of phenomena than any of the pure versions of TT or ST, this "eclectic" theory has remained highly speculative.

2.6. Concluding remarks

Although in the first years of intensive research on mindreading, the discussion evolved between different versions of TT, another player soon entered the game: ST. Instead of assuming that people rely on a rich body of law-like principles on how different mental states interact and link to behaviour, simulation theorists started to defend the idea that people use their own mental mechanisms as models for simulating the mental processes of other individuals. For some years, the participants of the mindreading debate defended TT over ST or the other way around. More recently, hybrid accounts have started to prevail. After Nichols and Stich (2003) and Goldman (2006) came out with their hybrid

accounts, the state of the art of the mindreading debate can be portrayed in the words of Hutto (2009, p. 223):

> After years of marshalling philosophical considerations and evidence from psychology, few TT/ST purists remain; most researchers working on everyday psychology accept that the best hope of providing an adequate account of the relevant abilities will draw on resources from both camps.

Thus the main question between defenders of TT and ST is no longer whether TT or ST is correct but rather how elements of TT and ST are best combined to explain human mindreading abilities. However, for the past decade or so, TT and ST haven't been "the only games in town." There is a new wave of theoretical approaches that are extremely critical of the whole framework that sees mindreading as the central ability in human social cognition. I will next turn to these new approaches and to how the standard theories of mindreading have been criticized.

Theories of mindreading, either in the form of TT, ST or some combination of them, provide possible explanations for the basic mechanisms of self- and other-directed mental-state ascriptions and discuss how mindreading typically develops in childhood. They also offer possible explanations for various break-downs of mindreading. All of this is done by appealing to empirical data that originates mostly from lab experiments, such as false belief tasks. However, as I will show in the next chapter, when theories of mindreading are construed as not only aiming to explain mindreading *per se* but as theories of human social cognition in general, they have serious limitations. The proponents of the standard theories of mindreading have taken for granted that mindreading is at the core of all human social cognitive abilities but this assumption has never been empirically tested. As the next chapter will make apparent, there are strong reasons to doubt that mindreading is ubiquitous in human social cognition.

3. AGAINST MINDREADING: THE INTERACTIONIST APPROACH

The 'theory of a theory of mind' is little more than a compound of philosophical misconceptions and inappropriate experiments.

Wes Sharrock and Jeff Coulter[29]

Although the main accounts of mindreading – TT and ST – began to develop as rivals, they share a whole set of fundamental assumptions and methodological strategies. The agreement of TT and ST on basic issues concerning social cognition has made their union in the form of recent hybrid accounts possible (Hutto, 2009, p. 223). A deep-seated assumption of the mindreading approach is that we understand other people primarily in terms of their mental states: both TT and ST consider mindreading to be the key to human social cognition. Articles on mindreading frequently begin with a statement about the centrality of mindreading in human social cognition. For example, Frith and Happé (1999, p. 83) claim that mindreading "appears to be a prerequisite for normal social interaction: in everyday life we make sense of each other's behaviour by appeal to a belief-desire psychology." Currie and Sterelny (2000, p. 145) argue: "mind-reading and the capacity to negotiate the social world are not the same thing, but the former seems to be necessary for the latter." Cruz and Gordon (2002, p. 9) maintain that mindreading "makes possible the rich social dynamic that pervades human life." Tager-Flusberg (2005, p. 276) writes: "successful social interactions depend on the ability to understand other people's behavior in terms of their mental states, such as beliefs, desires, knowledge, and intentions." The list could go on.

Recent years have witnessed growing resistance to the standard theories of mindreading: the importance of mindreading for human social cognition is being called into question, predominantly (but not exclusively) by philosophers with a background in the phenomenological tradition and by proponents of embodied and enactive approaches to cognition. TT and ST are heavily criticized by authors such as Shaun Gallagher (2001, 2004, 2007, 2008a–e), Daniel D. Hutto (2004, 2008a–b, 2009), Vasudevi Reddy (1996, 2008, Reddy & Morris, 2004), Ivan Leudar and Alan Costall (2009a), Hanne De Jaegher and Ezequiel Di Paolo (De Jaegher, 2009a, 2009b, De Jaegher & Di Paolo, 2007, De Jaegher, Di Paolo & Gallagher, 2010), Matthew Ratcliffe (2006, 2007), and Dan Zahavi (2007, 2008, 2011). In the following, I introduce the recent critique of the mindreading paradigm and sketch the main ideas of an alternative approach

[29] Sharrock and Coulter (2004, p. 580).

outlined by the critics. I will focus in particular on authors who argue that *engaged second person interactions* (carried out without the help of mind-reading) are more fundamental in human social cognition than observational third person mindreading. I will use the term *interactionism*[30] as a generic label for their approach. In the following six subsections, I will try to cover the central topics of the emerging interactionist framework.

3.1. Changing the explanandum

Interactionists argue that the mindreading approach is built on certain disputable philosophical assumptions that have been taken for granted and have not been critically investigated by researchers working within the mindreading paradigm. Thus the battle against the mindreading approach is largely a philosophical one. Proponents of theories of mindreading as well as interactionists agree that the general aim of studying human social cognition is to find out what enables people to understand and to interact with each-other. TT and ST imply that human social cognition relies on mindreading and explain mindreading as based on theorising and/or simulation. Interactionists, in contrast, argue not only against the view that mindreading relies on theorising or simulation[31], they are also strongly opposed to the idea that research on social cognition should take mindreading as its main focus of study. For example, Ratcliffe (2007, p. 234) writes: "Any enquiry into the nature of interpersonal understanding ought to begin with an adequate explanandum, an account of what is distinctive and central to a typical understanding of *people*." According to the view held by interactionists, mindreading is *not* an explanandum that is adequate for beginning the study of social cognition. They insist that instead of asking how people read minds, research on social cognition should begin by looking into how people interact in everyday life and explore what enables them to do so:

[30] I acknowledge that there are differences between the positive proposals of the critics of the mindreading approach, but I will here focus on what they have in common: the conviction that the mindreading paradigm is seriously flawed. The positive accounts of the critics are still in an early state of development and constitute a preliminary general framework rather than a set of particular theories. See Michael (2011), Bohl & van den Bos (2012), Herschbach (2012), and Overgaard & Michael (2013) for a critical discussion of interactionism.

[31] According to the thesis of *Direct Social Perception* (see e.g. Gallagher, 2008b) that many interactionists consider to be a part of their positive account, some mental states are directly perceivable, and thus theorising or simulation are considered to be unnecessary for understanding behaviour as psychologically meaningful. As far as mental states that go beyond what can be directly perceived (even by the standards of the DSP thesis) are concerned, Shaun Gallagher has joined forces with Daniel Hutto, who argues for the Narrative Practice Hypothesis (see e.g. Hutto, 2008, Gallagher & Hutto, 2008, and Gallagher & Zahavi, 2008). In a nutshell, the idea behind NPH is that children acquire a folk psychological competence as they are familiarized with certain types of narratives. NPH is, however, supported by only a few authors and as it is not central to interactionism as such, I will not dedicate much space to it here.

"investigating interaction is central to understanding social cognition." (De Jaegher, Di Paolo & Gallagher, 2010, p. 441)

In the standard mindreading paradigm, it is common to equate social cognition with an ability to *explain* and *predict* other individuals' behaviour via mindreading. However, it is plausible that mindreading is not the only, nor the primary, means that enables people to achieve explanation and prediction of other people's actions: people may anticipate and make sense of other people's behaviour by means of social norms, social roles, character traits and various other characteristics that fall out of the scope of mindreading (Andrews, 2009, 2012). Furthermore, to understand another person to whom you personally relate is not reducible to explaining or predicting her behaviour from a third person point of view (Ratcliffe, 2007, p. 25). Interactionists conceptualize social understanding as "a pragmatic ability to act appropriately in a particular situation" (De Jaegher, Di Paolo & Gallagher, 2010, p. 442) rather than as a theoretical ability to explain and predict others' behaviour via mindreading.

Interactionists emphasize that mentalistic explanations and predictions do not exhaust our social competence and are less central to social interaction than is assumed by TT and ST (De Jaegher & Di Paolo, 2007, p. 486). According to the critics, social cognition is primarily a matter of *mutual responsiveness* in everyday social practices rather than a matter of figuring out what is going on in another person's mind (see e.g. Ratcliffe, 2007, p. 49). Authors with a background in enactivism have proposed, for example, that instead of mindreading, social interactions are based on an ability to perceive and act upon *social affordances*[32] (Froese & Gallagher, 2012, p. 445). Interactionists characterize social cognition as a practical "know-how" (as opposed to theoretical "know-what" or "know-why") that enables us to initiate and sustain social interactions and to coordinate behaviour in various social situations. They point out that theories of mindreading hardly thematize, let alone explain, the issue of how people manage to appropriately respond to other people in particular social contexts. Interactionists point out that the limits of the standard mindreading approach become even more apparent when one considers breakdowns in social understanding: social understanding often fails because there is a lack of affective connection or engagement, not merely because people fail in their attributions of mental states (Ratcliffe, 2007, p. 174).[33] Some authors point out that social understanding includes not only an ability to understand others, but also "an ability to understand *with* others" (De Jaegher, Di Paolo & Gallagher,

[32] The term "affordance" was coined by James J. Gibson (1977, 1979). Affordances are properties of objects that allow inidividuals to perform certain actions. These properties are different for different individuals: for example, a chair allows a human to sit or to stand on it but it allows a cat to sleep on it or to use it as a means to get on the table. *Social* affordances are possibilities for *inter*action between two or more individuals.

[33] As I argue in appendix 3, mind*mis*reading may actually cause serious interpersonal problems and it is characteristic of certain psychopathologies, such as borderline personality disorder and schizophrenia. However, I agree that breakdowns of social understanding typically involve an affective component that theories of mindreading tend to ignore.

2010, p. 442, see also Gallagher, 2009).[34] In social situations, a common understanding of many aspects of the world is taken for granted but theories of mindreading arguably ignore it.

In short, interactionists argue that research on social cognition should first and foremost concentrate on the embodied and practical know-how that underlies real life social interactions instead of focusing on the issue of how individuals predict and explain other individual's behaviour by representing their mental states. The critique of the mindreading paradigm is not just a matter of philosophy – changes in the philosophical theory necessarily lead to changes in research methodology. An important point of controversy concerns the issue of what should be seen as the main unit of analysis in research on social cognition. For theories of mindreading, the basic unit is an individual mind figuring out other individuals' mental states. Interactionists are against such methodological individualism. Froese and Gallagher (2012, p. 437) even accuse theories of mindreading of "neuro-reductionism" – the view that the study of social cognition is exhausted by the study of sub-personal processes. The critics of theories of mindreading argue that the methodologically individualistic and neuro-reductive approach looses grip on what is essential to *social* cognition. *Inter*personal processes are claimed to be exactly that – processes between persons, not just representations locked inside people's heads (Reddy, 1996, Reddy & Morris, 2004, p. 653). Some interactionists argue that mental states can be literally shared among two or more people and are thus not necessarily owned by single individuals: "many thoughts, interpretations and viewpoints that emerge through conversation owe their creation to several people. They belong to nobody in particular and are shared products of interaction." (Ratcliffe, 2007, p. 174, see also Gallagher, 2008a, p. 555, and Krueger, 2013) Interactionists insist that social cognition cannot be studied without looking at situated social interactions between two or more embodied persons. They argue that a basic unit of analysis must minimally include two interacting subjects.

The interactionist perspective feeds the overall criticism of the *methodological limitations* of the mindreading paradigm (see e.g. Ratcliffe 2007, pp. 54–55 and 104–105). The main point is captured in the following quotation from Leudar and Costall (2009b, p. 11):

> … most of the Theory of Mind experiments substitute abstract puzzle-solving for situated, collaborative and embodied management of intentionality, with the result that they never really investigate the phenomena they are ultimately meant to explain.

Most mindreading experiments indeed investigate how children or adults attribute mental states to strangers (or puppets or cartoon characters) from a detached, third person point of view by passively observing a social scene. The

[34] Some advocates of theories of mindreading, for example Jane Heal, acknowledge that social understanding is not just about understanding the mental states of the other person, it is also an understanding of the world as a common ground between people.

critics argue that such studies reveal little, if anything, about the way social cognition functions in real life engaged social interactions: "it is not clear that false belief tasks even require the same kind of cognitive performance as most interpersonal understanding and interaction" (Ratcliffe, 2007, p. 105). The standard mindreading tasks are designed to identify the necessary elements of mindreading and to ascertain at what age children are able to attribute certain types of mental states. However, they provide no knowledge of how much children actually rely on mindreading in everyday social interactions. When and why people turn to mindreading in real life social situations has remained unstudied, presumably because of the overall assumption that mindreading is ubiquitous in everyday social cognition (Ratcliffe, 2007, p. 4). Interactionists, in contrast, claim that people in real life *never* rely on mindreading the way TT and ST assume[35] (Ratcliffe, 2007), or that people rely on mindreading *only in exceptional cases*, for example when the other's behaviour appears puzzling (Gallagher, 2001, p. 85).

3.2. The chicken-and-egg problem: Mindreading or social interaction?

When we shift the research focus from studying mindreading mechanisms to studying interactions between people, a question arises: is mindreading necessary for social interaction, as TT and ST seem to assume? Interactionists contend that it is not. The most radical critics of theories of mindreading (e.g. Ratcliffe, 2007) deny that mindreading has any psychological reality. Others concede that people sometimes attribute mental states but turn the claim that social interaction relies upon mindreading on its head, arguing that the development of mindreading *presupposes* basic non-mentalistic interpersonal abilities that are accomplished in embodied, emotional, and perceptual practices (see e.g. Gallagher & Zahavi, 2008, p. 187). Reddy and Morris (2004, p. 660), for example, argue against TT as follows:

> Theorizing presupposes the knowledge of others that is evident in engagement. Reflections upon, and theories about, other people's intentions and motivations do enter into everyday discourse, but these are developmentally and experientially secondary to actual engagement with these intentions and motivations. The Theory-Theory has simply not taken early development and engagement seriously enough.

Interactionists point out that young infants, adults with severe learning disabilities, and even pets interact with people, although they lack meta-representational abilities (see e.g. Leudar and Costall, 2009b, p. 7). For instance, three-year old children fail to demonstrate a capacity to attribute representational

[35] Ratcliffe (2007) uses the term "folk psychology" roughly in the same sense as I use "mindreading" here.

mental states in standard false belief tasks, but they nevertheless successfully interact with the experimenters (Gallagher, 2008e, p. 441). Already very young infants actively participate in exchanges of emotional expressions with their caregivers (Reddy, 2008) and immediately notice if the contingency of smooth social coordination is disrupted, as demonstrated for example in the "still face" experiments (Weinberg & Tronick, 1996). Infants not only respond to social stimuli, but also initiate and actively contribute to sustaining social interaction: when a caregiver suddenly stops communicating and displays a still face over a period of time, infants typically try to "repair" the communication by actively vocalizing and gesturing, and become upset if the partner remains unresponsive. It has been shown that already 6-week-old infants are sensitive to contingencies in social coordination: they enjoy live social interaction with their mothers via a double video link, but become distracted and disturbed after the online stream is seamlessly replaced by a replay of their mother's earlier communicative actions (Murray & Trevarthen, 1985). Infants quickly regain attention and positive affect after a second seamless shift to online interaction (Nadel et al., 2009). The general point of the interactionists is that in most social situations, people need to be able to socially interact with others and that for this purpose, mindreading is often unnecessary. In short, they argue that the proponents of theories of mindreading are simply wrong in assuming that social interaction presupposes mindreading.

The chicken-and-egg problem of whether mindreading precedes social interaction or vice versa is actually not as black and white as it may seem from the way interactionists present it. Most proponents of the mindreading paradigm explicitly acknowledge that the development of mindreading proper relies on other, more basic abilities: for example, in Baron-Cohen's theory, the Mindreading Mechanism can only develop if the Intentionality Detector, Eye Direction Detector and the Shared Attention Mechanism function properly (Baron-Cohen, 1997; see section 2.3.2 of the current summary article). Furthermore, Baron Cohen and Cross (1992) have explicitly argued for the importance of perception in social cognition. Advocates of the mindreading approach are unlikely to question the claim that some forms of social interaction are possible without mindreading – after all, many non-human social species interact without displaying any evidence that they are able to attribute mental states. What the proponents of the mindreading approach do assume is that mindreading is necessary for *full-blown human social competence*. It is assumed that when children begin to mindread, they acquire more sophisticated and more efficient means for social understanding and interaction (Spaulding, 2010, pp. 125–128). Early non-mentalistic abilities for social interaction are considered to be necessary but insufficient for full-blown human social cognition; they are seen as mere precursors[36] to mature social cognitive abilities. According to inter-

[36] Gallagher (2008c, p. 165) defines 'precursor' as an ability that disappears when the 'real' capability – mindreading – takes over in course of development. He argues that non-mentalistic abilities are not precursors to mindreading in this sense. This use of the term in

actionists, however, non-mentalistic social abilities remain central to social cognition throughout life, and in most social situations, social interaction and social understanding is arguably achieved without the help of mindreading (Gallagher, 2008c). This brings me to the issue that the negative claims of interactionists about the insignificance of mindreading are actually as speculative as the claim that mindreading is ubiquitous in human social cognition. Interactionists correctly point out that in the framework of theories of mindreading, embodied non-mentalistic abilities for social interaction have not been seriously investigated. However, as I argue in appendices 1 and 3, the exact role of mindreading in human social cognition needs further empirical investigation. In appendix 2 I show that the critique of the mindreading approach is partly targeted at a straw man rather than at theories of mindreading properly understood.

One argument that supports the view that people may mindread rarely rather than all the time, is that mindreading may be too complex to be an efficient tool for social interaction. Several philosophers from the analytic tradition have argued that mindreading is cognitively too demanding to enable the swift and effortless social coordination that we typically experience most of the time while socializing (see e.g. Morton, 1996, 2003, Bermudez, 2003; for counter-arguments, see Spaulding, 2010). Morton (1996) highlights three basic problems with mindreading: *holism*, *entanglement*, and *complexity*. Holism is a serious problem especially for TT, but potentially also for ST: it is virtually impossible to judge which mental states underlie any particular action when "any belief or any desire can be rationally consistent with any action, given suitable other beliefs and desires" (Morton, 1996, p. 128)[37]. Entanglement means that many of our decisions are related to outcomes that depend on the decisions of other people; it means that any decision on the part of one person changes the information on which the decisions of everyone else is based. If social interaction were based on mindreading, choosing what to do in a social situation would become an enormously complex task, especially when more than two people are involved in an interaction. Morton (1996, 2003) proposes that instead of mindreading, people base most of their social decisions and predictions on the assumption that others usually act according to shared norms. Likewise, Bermúdez (2003, p. 47) argues that mindreading in multi-agent interactions leads easily to combinatorial explosion and computational intractability, and maintains that we turn to mindreading "not as a mainstray of our social understanding, but rather as the last resort." However, even if these authors are right that mindreading is too complex to be a necessary component of all social interaction, and if it is true that people mindread only in rare cases, a question still remains: what kind of cases are they and why has the ability to attribute

---

the context of social cognition is, however, misleading: no one has defended the view that once children become capable of mindreading, their non-mentalistic social abilities cease to function. For example, in Baron-Cohen's modular theory, the three modules (ID, EDD, and SAM) mature before ToMM, and continue to provide input for ToMM throughout life.

[37] See also my discussion of the frame problem in 2.3.1.

mental states evolved in the first place? In other words: what is the function of mindreading? Neither theories of mindreading nor the interactionist accounts provide a satisfactory answer to this question.

3.3. Embodied social cognition

According to interactionists, many flaws of the mindreading approach stem from a presupposed "disembodied" cognitivist framework which implies a gap between the mental and the physical (Reddy, 2008, Leudar & Costall, 2009a, Sharrock & Coulter, 2004). Interactionists often begin their critique of the mindreading approach by arguing against the view that mental states, unlike behaviour, are unobservable. Allegedly, a fundamental presupposition underlying theories of mindreading is the assumption that *other minds are completely unobservable*, which erroneously motivates the postulation of special mindreading mechanisms, such as theorising or simulation (c.f. Gallagher & Zahavi, 2008, Gallagher, 2004, 2008b, Zahavi & Gallagher, 2008, Zahavi, 2007, 2005, Zahavi & Parnas, 2003, Leudar & Costall, 2009b, Hutto, 2009, Reddy, 2008, Ratcliffe, 2007). According to intractionists, TT and ST deny that we can ever truly *experience* other people's mentality (and in the case of TT, even our own mentality) – other minds remain mere chimera that we need to infer or imagine, but can never truly meet. Interactionists propose that instead of the Neo-Cartesian assumption that postulates a gap between the mental and the physical, and thereby also between any two minds, a more adequate starting point would be to assume that "human beings are primordially connected in their subjectivity." (Zlatev et al., 2008, p. 3)

On the basis of phenomenological arguments and enactive and embodied accounts of cognition, a sharp distinction between mental and bodily states is denied by interactionists. Bodily expressions are taken to be parts of mental states. If so, this would imply that mental states are sometimes directly perceived, rather than merely inferred:

> Gestures, expressions and actions are not just external expressions of internal thought processes; they are part of those processes. Hence we really can perceive, to some extent, the mental lives of others. (Ratcliffe, 2007, p. 148).

This view is known as the thesis of *Direct Social Perception* (DSP), and many interactionists consider it to be a part of their positive account. According to DSP, it is possible to *perceive* (some of) other people's mental states because mental states are embodied in expressive behaviour, and because perception is a relatively "smart" process, able to deliver not only thin sensory information about physical behaviour, but also the thick psychological meaning of it (Gallagher, 2008b). In the interactionist framework, mirror neurons and other neural resonance processes are taken to be a part of the mechanism of social perception rather than processes that rely on simulation apart from perception (Gallagher, 2007). Behaviourism is carefully avoided by emphasizing that

mental states are not reducible to observable behaviour: mental states are said to transcend their bodily expressions (Zahavi, 2007), but behavioural expression is nevertheless conceptualized as being literally a part of mental states. DSP is supported by empirical evidence which indicates that bodily expression is experienced as an essential aspect of certain mental states. For example, people with Moebius syndrome[38], who are unable to express emotions facially, tend to have diminished experience of affect (Cole, 2010, Krueger & Overgaard, 2012). Various studies suggest that gestures serve as thinking and memorizing aids (see the review of Goldin-Meadow, 1999, and Ratcliffe, 2007, pp. 147–148).

Unfortunately, the claim that theories of mindreading assume that we can only perceive meaningless physical behaviour and need to add an extra layer of psychological interpretation on top of it, is a straw man. Theories of mind-reading indeed tend to underestimate the role of embodiment in social cognition, but as I will show in appendix 2, the interactionist attack against the alleged unobservability assumption of theories of mindreading is misplaced and confuses the discussion. Proponents of theories of mindreading do *not* deny that people sometimes have an experience of simply seeing another person as being in a certain mental state, for example, as happy, bored, or worried. Nobody questions the view that actions are typically perceived as psychologically meaningful ("he waves goodbye"), rather than as sequences of mechanical bodily movements in need of interpretation ("he raises his upper limb and moves it back and forth in the air"). The real controversy between the mind-reading accounts and the interactionist view of direct social perception lies in how the sub-personal processes that underlie mindreading, including the experiences of directly apprehending another person's mental states, are conceptualized, not in whether such experiences exist.

As far as embodiment is concerned, a shortcoming of the accounts of mind-reading that is more serious than their alleged adoption of an unobservability assumption is their insensitivity to the role of *bodily responsiveness* in social cognition. Neonate imitation studies (Meltzoff & Moore, 1977, 1983) demonstrate that newborns are already responsive to certain facial expressions. Interactionists argue that neonate imitation occurs not because, as theory-theorists suggest, infants are able to theorise as soon as they are born (see my discussion of Gopnik's view that suggests this in section 2.3.1.), but because there is an inborn cross-modal link between the structure of one's own body and the bodies of other people at the level of perception, affect and proprioception (Gallagher & Meltzoff, 1996[39], Gallagher, 2004). Even low-level simulation accounts that pay more attention to affective and sensory-motor processes than TT (as for example in Goldman, 2006) fail to do justice to the role of bodily responsiveness (Gallagher, 2008c, p. 170). When I interact with you, various

[38] Moebious syndrome is a congenital neurological disorder which is characterized by facial paralysis.

[39] Andrew Meltzoff has co-authored with both, Alison Gopnik (Gopnik & Meltzoff, 1997), and Shaun Gallagher (Gallagher & Meltzoff, 1996). He seems to support different inter-pretations of the imitation studies, depending on whom he is writing with.

affective and sensory-motor processes arise in my body and directly feed into my experience, affect my behaviour and inform my understanding of you and of the whole situation, thus influencing the whole social interaction. ST of low-level mindreading focuses on explaining how one recognizes emotions and motor intentions in another person by registering their facial expressions and bodily gestures, but the issue of how one *reacts* to perceived mental states, and how the reaction forms a part of the understanding of the situation, is not raised (see also Bohl & van den Bos, 2012, p. 4). For example, neither TT nor ST discuss how a person's affective response (e.g. fear) to a partner's perceived mental state (e.g. anger) affects social interaction.

3.4. Second person versus third person mode of social cognition

It has become a part of the interactionist agenda to emphasize the prevalence of the *second person* mode of social cognition in contrast to the *third person* mode (see Ratciffe 2007, pp. 152–155, Reddy, 2008, pp. 26–42, Reddy & Morris, 2004, Gallagher 2001, 2007, 2009). The distinction between the second person and the third person modes of interpersonal understanding derives from respective grammatical categories expressed by personal pronouns. Interactionists have more than just a grammatical artefact in mind, however: by differentiating between the third person and the second person modes of social cognition, they aim to highlight the distinction between observing others from a detached point of view and being engaged in reciprocal social interaction. Unlike the third person (*he*/*she*), the second person (*you*) exists only in relation to *I*: it emerges when one directly addresses and/or is being addressed by another person. Drawing on Martin Buber's "Ich und Du" (1923) and the works of classical phenomenologists such as Martin Heidegger, Aron Gurwitsch and Alfred Schütz (see Ratcliffe, 2007), interactionists argue that observational third person understanding of others develops against a backdrop of second person interactions and thus that the second person mode of social cognition is more fundamental than the third person mode (Gallagher, 2001, De Jaegher & Di Paolo, 2007, p. 503).

Developmental psychologist Vasudevi Reddy (1996) was the first to point out (in the context of mindreading) that people primarily understand each-other using "second person information." For Reddy, second person information is something that becomes evident in interpersonal *emotional engagement*[40], giving access to aspects of another person's intentionality that otherwise remain inaccessible (Reddy, 1996, p. 140). In Reddy's example, seeing a person smiling

[40] For Reddy, *emotional engagement* refers to mutual emotional responsiveness that arises in social interactions. Other authors also talk about engagement, but they may have something different in mind. For example, De Jaegher, Di Paolo and Gallagher (2010, p. 442) define *engagement* as capturing "the qualitative aspect of social interaction once it starts to 'take over' and acquires a momentum of its own."

at a third person and seeing a person smiling at you are qualitatively different experiences: In the latter case, what you see of the person and what you feel in response become intermingled as aspects of the same experience.

> Within active emotional engagement your perception of the other always involves proprioceptive experience of self-feelings-for-the-other, and your proprioception of the self always involves perception of other-feelings-for-self. (Reddy 2008, p. 30).

When someone smiles at you, you are called upon to respond (even not responding will count as a response), which is not the case when you see somebody smiling at a third person.

The importance of the "second personal" way of understanding others is emphasized in various expressions like "second person information" (Reddy, 1996), "second-person interaction" (Gallagher, 2001, 2004, 2007, 2009, Ratcliffe, 2006), "second-person relationship" (Gallagher, 2004), "second person(al) contexts" (Hutto, 2004; De Bruin & Strijbos, 2010), "second-person stance" (Ratcliffe, 2006), "second-person perspective" (Gallagher, 2009, Reddy, 2008) etc. Reddy (2008), emphasizing the role of second person information in social cognition, even calls her account of social cognition "a second-person approach" and argues that research of social cognition needs to turn to "second-person methodology" (p. 32–39) , i.e. to participatory methods. On a closer look, speaking about the "second person" turns out to be somewhat confusing, however, because the expression as such does not refer to any clear and well defined category and there is a lack of consensus on how to define it. Different authors focus on different aspects of social cognition when they talk about the "second person" in contrast to the "third person." For example, Reddy highlights the difference between being emotionally engaged and being emotionally detached, whereas Gallagher accentuates the difference between interacting with others and merely observing or thinking about others. Emotional engagement and social interaction often go together, but they may also come apart: sometimes observing another person or merely thinking about someone triggers a strong emotional response towards that person in the absence of any social interaction, whereas social interactions may take place in the absence of emotional engagement.

One option is to go back to grammatical categories and to assume that whatever a second person mode of social cognition includes, at its core is an understanding of another person as a *you*. Ratcliffe (2007, p. 152) explicitly defines "the distinction between second-person and third-person understanding" as "the difference between understanding someone as a 'you' and understanding someone as a 'he' or 'she'." But in this case, it is not obvious in what sense the second person understanding is more fundamental than the third person understanding, as interactionists insist. From an early age, we not only interact with others from a second person perspective, we also observe them from a third person point of view. So both perspectives seem to be equally present in early

childhood. Later in life, social interactions involve not only dyadic (I-you) interactions, but also triadic (I-you-he/she) interactions (Cleret de Langavant et al, 2013). Our real life social encounters constantly shift between those that can be described by using a second person pronoun and those that can be described by using a third person pronoun. Furthermore, if we draw on grammatical categories, we cannot avoid the fact that having three person categories is not universal across languages: "many non-Indo-European languages have four person categories, not three, adding a 'first-person inclusive' to denote the union of speaker and addressee" (Evans, 2013). Unlike in Western individualist societies, in collectivist cultures in Asia and other parts of the world, the first-person inclusive may be seen as the most fundamental social category rather than the second person, which suggests that people may experience daily social encounters quite differently in different cultures.

Presumably, what interactionists have in mind when they talk about the fundamentality of the "second person" in social cognition is something more primordial and universal than the grammatical categories – something that exists before children have acquired language and learned to experience the world through linguistic categories. Thus it might be more adequate to understand the "second person" as a metaphor that has come to be used as a shorthand for talking about some basic aspects of interpersonal understanding that the critics of the mindreading account wish to emphasize. It is used to point out that *social interaction* and *emotional engagement* are two important (although not necessarily always co-occurring) components of social cognition that any comprehensive account of social cognition needs to explain. The ambiguity in the use of the term "second person" has motivated attempts to find ways to formulate some crucial aspects of social cognition more clearly. For example, Monika Dullstein (2012), drawing on ideas from Stanley Cavell (1976), proposes that the difference between third person and second person understanding should be conceptualized as a distinction between *knowing* others (having justified true beliefs about other minds) and *acknowledging* others (responding to others appropriately). Recently, under the heading of the "second person," neuroscientists have also started to look for new conceptual and methodological tools to study social cognition as an interactive process. For example, Schilbach et al. (2013) recently published a target article in *Behavioural and Brain Sciences* entitled "Towards a second person neuroscience," where they propose to operationalize the "second person" by using interaction and emotional engagement as two parameters that can be manipulated in an experimental design.

3.5. Social interaction as constitutive of social cognition

Interactionists emphasize that it is important to study people while they interact with others in real life rather than while they merely observe others in impoverished lab conditions, because different cognitive processes are involved in

the two conditions. They expect that research on social interaction will reveal that while interacting, people do not actually rely on mindreading, but instead make use of perceptual and sensory-motor abilities and non-mentalistic understanding of shared contexts.

Interactionists argue that a further reason to focus on social interactions is that social interaction is more than simply a sum of individual processes; it is "not reducible to the workings of individual cognitive mechanisms" (De Jaegher, Di Paolo & Gallagher, 2010, p. 441). However, not any kind of action of two or more individuals sharing a physical space is an *inter*action – interaction involves mutual influence or *coupling*[41] between two or more agents. De Jaegher, Di Paolo and Gallagher (2010, pp. 442–443) propose the following definition for social interaction: "Two or more autonomous agents co-regulating their coupling with the effect that their autonomy is not destroyed and their relational dynamics acquire an autonomy of their own." In other words, interaction of two or more agents which acquires its own momentum, but does not destroy the autonomy of the interacting agents[42], is what interactionists consider *social* interaction to be. That interaction processes have an autonomy of their own is most clearly seen in cases wherein social interaction emerges or carries on despite the efforts of the participants to avoid or to terminate it. Here is a simple example by De Jaegher and Di Paolo (2007, p. 493):

> Consider the situation in a narrow corridor when two people walking in opposite directions have to get past each other. They have to decide whether to continue walking as they are, or shift their movement to the right or to the left. Occasionally, such encounters unfold like this. Instead of choosing complementary movements that would allow them to carry on walking, the individuals move into mirroring positions at the same time. This unintended coordinated change in individual position creates a symmetrical mirroring relation. This symmetry, in combination with the spatial constraints of the corridor, increases the likelihood that the next move will also be a mirroring one (there are not many other moves available). Thus, the coordination maintains a property of the relational dynamics that forces the individuals to keep facing each other and consequently to remain in interaction (in spite of, or rather because of, their efforts to break from this situation).

This example is meant to demonstrate that a system of two or more interacting social agents has special properties that are irreducible to properties of single individuals. According to the view of DeJaegher, Di Paolo and Gallagher, social interaction is not just a result of the social cognitive processes of the participating individuals, but it is partly sustained by the unfolding of the interaction process itself: "the agents sustain the encounter, and the encounter itself influen-

[41] *Coupling* is a central concept in the enactivist theory, which refers to one-sided or mutual influence between the parameters of two or more systems (see De Jaegher, Di Paolo & Gallagher, 2010, p. 441).

[42] "If one agent becomes the sole regulator of the coupling, as in the use of a tool, this is no longer social interaction." (ibid. p. 444)

ces the agents" (De Jaegher & Di Paolo, 2007, p. 492). In short, interactionsts argue that in some cases social interaction *enables*, and may even *constitute*[43] social cognition: "interactive processes are more than a context for social cognition: they can complement and even replace individual mechanisms." (De Jaegher, Di Paolo & Gallagher, 2010, p. 441)

Inspired by the ideas associated with extended cognition, interactionists argue that cognitive processes are not limited to what is happening inside the skull; they are partly offloaded onto the environment and to the interaction dynamics. In social interaction, *coordination* – a non-accidental correlation in the activity of the two or more systems[44] – plays an important role. Interactionists point out that coordination is widespread not only in social systems, but also in many physical and biological systems – for example, pendulum clocks synchronise over time when placed in each other's vicinity (De Jaegher & Di Paolo, 2007, p. 490). "If pendulum clocks can do it without mechanisms for "timing the beat" and "forming a temporal estimate", why can't babies?" (ibid. p. 499), they ask. Since many non-social systems achieve coordination via physical and biological mechanisms without having any special mechanisms for cognitively representing the properties of the world, it is possible that human social coordination also makes direct use of such physical and biological mechanisms so that aspects of the interaction dynamics themselves, not just the individual cognitive processes, allow the interaction to be sustained. Thereby they appeal to parsimony: "Why develop a complicated internal capacity when the environment can do the job for you?" (Ratcliffe, 2007, p. 108)

In order to better understand the interactionist point about social interaction enabling or constituting social cognition, let me come back to the experiment involving young infants who interact with their mothers via a double video monitor (Murray & Trevarthen, 1985). Why does the behaviour of the infant change when the live video stream of their mother is replaced by an earlier recording? An individualist explanation appeals to special cognitive mechanisms that infants presumably have for detecting contingency and for noticing the lack of contingency, in which case their behaviour changes. An interactionist explanation gives the interaction process an enabling or a constitutive role: an infant's behaviour changes because the properties of the system of one-sided coordination is much less stable and more difficult to sustain than reciprocal coordination (see De Jaegher & Di Paolo, 2007, pp. 489–490, and De Jaegher, Di Paolo & Gallagher, 2010, p. 441).

---

[43] De Jaegher, Di Paolo & Gallagher (2010, p. 443) distinguish between contextual, enabling, and constitutive factors and define them as follows: "F is a contextual factor if variations in F produce variations in X", "C is an enabling condition if the absence of C prevents X from occurring" and "P isa constitutive element if P is part of the processes that produce X."

[44] De Jaegher, Di Paoplo and Gallagher (2010, p. 441) define coordination as "non-accidental correlation in the activity of two or more systems that are coupled at present or were coupled in the past, or are or were coupled to another system in common, over and above what is expected from their normal behaviour in the absence of such couplings."

3.6. Contextual understanding

In most social situations, other people appear to us primarily as agents whose actions are framed in their practical activities and the meaning of their behaviour is not only expressed in embodied actions, gestures and facial expressions, it is also supported by the wider context of the situation. This means that in order to have an optimal understanding of other people's actions and in order to be able to act appropriately in social situations, one needs to have an adequate understanding of the context of the situation.

Interactionists, especially Shaun Gallagher (see e.g. Gallagher, 2004, 2005, 2008c, 2012, Gallagher & Zahavi, 2008, Gallagher & Hutto, 2008, Froese & Gallagher, 2012), have been inspired by Colwyn Trevarthen's notions of *primary intersubjectivity* (Trevarthen, 1979) and *secondary intersubjectivity* (Trevarthen & Hubley, 1978). Primary intersubjectivity, already manifest in the behaviour of the newborn, refers to early, embodied capacities for "face-to-face" interaction that rely upon direct perception of emotions and intentions, as well as upon bodily and affective responsiveness. At around 12 months of age, children enter into *secondary intersubjectivity*: they start to display abilities for *shared attention* and begin to tie actions to *pragmatic contexts*. According to Gallagher (2008c), primary and secondary intersubjectivity remain at the core of our social competence throughout life and constitute the non-mentalistic basis for human social cognition.

Interactionists criticize theories of mindreading for not providing any comprehensive explanation of the role of contextual understanding of social situations, and for putting the whole emphasis on attributions of mental states as causes or reasons for actions (see e.g. Ratcliffe, 2007, p. 150). In contrast, interactionists insist that usually people are not trying to get into other people's minds, but rather into their worlds by entering into and creating shared practical contexts (Gallagher, 2007, p. 354). According to Ratcliffe (2007, p. 150), we interpret actions "by progressively adding layers of surrounding situational context and moving 'outward', rather than by moving 'inward' and postulating internal states as causes." Interactionists also hypothesise that serious problems with intersubjective understanding, such as in people with autism, involve difficulties with contextual understanding rather than problems with mental state attributions (Gallagher, 2001). But how does contextual understanding work in healthy subjects?

Gallagher and Zahavi analyse the importance of practical contexts by drawing on the classical phenomenological descriptions of the life-world by Husserl, Heidegger, Gurwitsch and other phenomenologists (Zahavi, 2005, Gallagher, 2012, Gallagher & Zahavi, 2008). With Hutto they argue that in the process of becoming language users, children are immersed in various *narrative practices* and thereby gradually acquire a more complex understanding of broader cultural and social contexts and learn to think about reasons for actions (see Hutto, 2008a–b, Gallagher & Hutto, 2008, Gallagher & Zahavi, 2008). Ratcliffe (2007, p. 58–120) emphasises that people *take for granted* that they

live in the same world with others and argues that people let the shared world do a lot of the work of facilitating social cognition. The shared world, however, is not just the physical world we live in – it is a world of *social norms* and *conventions*, which regulate social behaviour, arguably without mindreading (see Ratcliffe, 2007, pp. 58–60). In other words, it is not that mindreading enables us to figure out that our understanding of the world largely overlaps with how other people understand the world, but any act of mindreading already *presupposes* the existence of such common ground:

> An understanding that at least some aspects of a situation are shared is not assigned to others in the form of a belief system but presupposed. For example, when trying to pass a person in a busy shop, one might think 'he wants to get to the checkout' but one would not ordinarily think 'he believes he is in a shop'. (Ratcliffe, 2007, p. 57)

De Bruin & Strijbos (2010) similarly argue that in many social situations, the practical context of the situation is interpersonally shared, so that other people's actions appear as directly comprehensible and there is no need to draw on explicit or implicit inferences about the context in order to figure out their reasons for action. They hypothesise that when another person's reasons for action remain unclear and it is important to get them right, people typically engage in *reason conversations*: they ask for reasons either from the agent or from other people who may know better. De Bruin and Strijbos propose that people only turn to mindreading, which is a highly fallible strategy, in cases where it is either unimportant to get the other person's reasons right, or when it is for some reason inappropriate or impossible to ask about them from the agent or from other people. They are also sympathetic to the idea that mindreading may be more adequately characterized as a *regulative*, rather than a *descriptive*, practice. The idea of mindreading being a regulative device has been elaborated by Mameli (2001), McGeer (2007), and Zawidzki (2008, 2013), whom I would classify as non-interactionist critics of the standard theories of mindreading. I discuss their ideas more thoroughly in appendix 3, where I hypothesise that mindreading may serve the purpose of regulating *relationships*.

To be fair, most theories of mindreading have tried to accommodate the existence of a large amount of shared beliefs about the world and acknowledge the need to include contextual information in the process of mindreading. For example, recall the argument that mindreading cannot be strongly modular, because contextual information needs to be part of the mindreading process (see section 2.3.2.). According to child-scientist versions of TT, folk-psychological principles depend on *ceteris paribus* clauses; i.e. it is assumed that in any particular act of mindreading, the context of the interpreted behaviour needs to be taken into account when selecting the appropriate principles for a particular situation (Bruin & Strijbos, 2010, p. 260). One of the biggest challenges for TT is to explain *how* folk-psychological principles are selected in any particular context – this issue is part of the notorious frame problem (see section 2.3.1.). According to ST, people assume, as a default strategy, that other people have the

same beliefs about the world as they do. According to the theory of Nichols and Stich (2003), at the beginning of the mindreading process, the mindreader includes all of her own beliefs in the Possible World Box. But simulation-based theories need to explain how the mindreading system calculates interpersonal discrepancies, and here contextual information comes into play. In cases of high-level mindreading, for example, the system needs to select appropriate input states for simulation and such selection mechanism needs to reckon with the context. In sum, in order to explain how mindreading works in different contexts, TT needs to explain how the system calculates the relevance of different folk-psychological principles and ST needs to explain how the system selects appropriate pretend states for simulation; in both cases, the problem of computational complexity is lurking behind the scenes (see De Bruin & Strijbos, 2010, p. 260).

Whereas theories of mindreading assume that our contextual knowledge of the shared world mainly functions as input for mindreading, interactionists argue that most of the time, contextual understanding of situations enables successful interaction with, and successful understanding of, others without mindreading. For instance, Ratcliffe (2007, p. 90) argues against ST that instead of predicting what another person will decide to do based on simulating what *I* would decide to do in her situation (or in Gordon's version of ST – what I would do if I *was* the other person), people are more likely to simply anticipate that other people typically behave according to shared norms. Moreover, if a person transgresses a social norm, she not only becomes incomprehensible to others, but a subject of criticism and, in more serious cases, social sanctions: "we make each other predictable by putting norms in the world for all to follow and also by interacting with each other in such a way as to influence behaviour so that it conforms to a greater degree with norms." (Ratcliffe, 2007, p. 18)

Interactionists are not the first to emphasize the role of social norms in social cognition. Several non-interactionist critics of the mainstream theories of mindreading, and even some proponents of the standard theories of mindreading themselves – such as Jane Heal – stress the importance of norms in social cognition. Whereas Heal (2003) mainly speaks about norms of rationality that govern our thinking and mindreading, Morton (1996) and Bermúdez (2003) propose that normative understanding of social situations is related to an understanding of the *social roles* of the given culture. In a typical restaurant scene, for example, there is usually no need to attribute mental states to the waiter to be able to understand what he is doing and why, insofar as he behaves according to his role (Bermúdez, 2003, pp. 43–44). It is part of the role of being a waiter to bring the menu to the table and to ask from the customers what they would like to have, just like it is a part of the role of the customer to order a meal and to pay for it before leaving the restaurant. Kristin Andrews (2009, 2012) claims that implicit knowledge of social norms of one's community is fundamental to social cognition and presupposed by mindreading. On her account, people mindread in order to explain behaviour that violates shared norms, which suggests that mindreading presupposes an understanding of social

norms. According to Zawidzki (2008, 2013), mindreading is only possible because various normative and educational social practices shape our minds and behaviour to be interpersonally comprehensible, and in his theory, mindreading itself is a specific tool for *mindshaping*.

In short, interactionists draw attention to the necessity of including the issue of contextual understanding in research on social cognition, but their explanations of the cognitive processes that enable contextual understanding are still in infancy: a comprehensive theory of contextual social cognition is yet to be worked out. At the same time, some non-interactionist critics of the standard theories of mindreading have proposed more detailed accounts of how social norms and roles support social cognition.

3.7. Concluding remarks

In sum, interactionists argue that mindreading is *not* what enables people to understand and interact with each other in most real life situations and that the mentalistic assumption is merely an unfortunate idea of philosophers and psychologists rooted in a certain philosophical background. The assumption that mindreading is the key to human social cognition has directed the way in which empirical research has been carried out in the standard mindreading framework. Social cognition of children and adults has been studied mainly in experimental conditions of low ecological validity, devoid of reciprocal social interaction and real life situational context. In such experiments, subjects have typically observed social scenes and have thereby been forced to mindread, because more natural and embodied ways of social understanding have been ruled out by the experimental design. This in turn has kept the underlying theoretical implication unchallenged: since the majority of empirical studies call for mindreading, it has been easy to continue thinking that mindreading is ubiquitous in social cognition.

In contrast, interactionists insist that instead of focusing on mindreading, research of social cognition needs to shift its focus to social interaction. They argue that for the most part, people interact with and understand each other on the basis of non-mentalistic interactive abilities, such as direct perception, bodily responsiveness, and non-mentalistic contextual understanding of shared situations. They also emphasize that social interaction is not just a result of social cognition, it in turn scaffolds social cognition, thus being an enabling, and possibly even a constitutive, factor of social cognition. In the interactionist framework, only a minimal role is given to mindreading. Interactionists typically claim that people mindread only in rare cases; for example when the other's behaviour appears particularly puzzling.

My main worry with interactionism is that in arguing that mindreading is peripheral to social cognition, interactionists have gone to the extreme and have thrown the baby out with the bath water. The claim that mindreading is absent in most social situations is actually just as speculative as the claim that

mindreading is ubiquitous in human social cognition: it is currently not supported by empirical evidence. The fact that theories of mindreading are based on studies in which people primarily observe others instead of interacting with them is not a reason to discredit theories of mindreading completely. It is undeniable that also in real life, people sometimes observe other people and think about other people's mental states. Theories of mindreading provide so far the most elaborate explanations of such cases of social cognition. If mindreading plays no important role in social cognition, as the interactionists claim, why has the ability to attribute mental states evolved at all? In fact, it has been shown that children pass interactive versions of false belief tasks earlier than the classical observational false belief tasks (Buttelmann, Carpenter, & Tomasello, 2009, Knudsen & Liszkowski, 2012), which suggests that interactive contexts may facilitate mindreading rather than prevent it. What needs to be studied is how non-mentalistic interactive abilities and mindreading both contribute to human social cognition. Then it will also be possible to spell out the specific role of mindreading that neither theories of mindreading nor interactionism have captured so far.

4. SUMMARY AND CONCLUSIONS

My dissertation contributes to the interdisciplinary study of human social cognition. The main aim of the thesis is to explain why we need to move beyond the current two-sided debate between *theories of mindreading* and *interactionism*, and to demonstrate how new research questions and hypotheses grow out of an integrative approach.

The dissertation consists of a general summary article and three theoretical research articles that have been published in international peer-reviewed journals (appendices 1–3). The summary article begins with a chapter on methodology, where I consider different ways of applying philosophical methods in the interdisciplinary study of social cognition. I discuss how philosophical competence can contribute to interdisciplinary research on social cognition in general, and I explain how it contributed to the research articles that form the main part of my dissertation. Most importantly, I show that philosophical competence is applicable in the critical evaluation of theoretical frameworks and methodologies of ongoing research paradigms. Training in philosophy is crucial in interdisciplinary research, as it fosters the asking of questions at the intersection of different disciplines and thereby contributes to the development of new hypotheses and research methods. At the same time, it facilitates systematic reflection upon how theoretical claims and empirical results of different disciplines and paradigms hang together or contradict one another, and enables to build novel and more comprehensive theoretical frameworks. Another area where philosophical competence is particularly helpful is conceptual clarification, especially in the form of the analysis of existing conceptual frameworks and in the process of "tailoring" new concepts for specific theoretical purposes. I also look at the potential role of phenomenology in the interdisciplinary research of social cognition and show that besides reinterpreting scientific findings in a phenomenological framework, or using phenomenological descriptions as *explananda* for social cognition research, it is possible to do neurophenomenological experiments on social cognition and to "frontload" phenomenological insights on social cognition into experimental design.

I begin chapter 2 by introducing early mindreading studies and continue by outlining the standard theories of mindreading that have dominated research on social cognition for more than three decades: *theory-theory, simulation theory*, and their hybrids. According to theory-theory, people make inferences about other people's mental states by using a tacit theory about how different types of mental state link to each other and to behaviour. The *child-scientist versions* of theory-theory explain the development of mindreading in children by analogy to theory-acquisition in science, whereas according to *modularist versions* of theory-theory it is a matter of the maturation of an innate mindreading module. According to simulation theory, people do not use a theory at all, but instead read each other's minds by mentally putting themselves into the shoes of other people: they imagine what it would be like to be in the other person's position, and attribute the mental states that arise in their mind to the target. More

recently, hybrid theories have emerged that combine elements of both theory-theory and simulation theory.

In chapter 3, I introduce a recent critical approach that challenges the standard theories of mindreading: *interactionism*. In general, interactionists insist that the research focus needs to shift from mindreading to social interactions, which is why I have adopted the label "interactionism." Interactionists strongly oppose the prevailing assumption that mindreading is the key component of human social cognition that enables people to understand and interact with each other throughout all or most social situations. In contrast, they argue that for the most part, people interact with and understand each other on the basis of non-mentalistic interactive abilities, such as direct perception, bodily responsiveness, and a non-mentalistic contextual understanding of shared situations. Interactionists give only a minimal role to mindreading in human social cognition: they argue that people rarely attribute mental states to others, presumably only when the other person's behaviour puzzles them. A frequent interactionist counterargument to theories of mindreading is the claim that the standard theories are based on the flawed assumption that other minds are unobservable. Interactionists also point out that theories of mindreading are based exclusively on lab experiments where test subjects are asked to passively observe social scenes from a detached third-person point of view and that thereby engaged second-person interactions (that are far more essential to social cognition) have been ignored. According to interactionists, not only are different individual cognitive processes involved in engaged social interactions (as compared to detached third-person observations), but also, the interaction process itself plays an enabling and possibly even a constitutive role in social cognition.

I fully agree with interactionists that the standard mindreading paradigm is limited and incapable of providing a complete account of the full spectrum of human social cognitive abilities. We certainly need to study how social interactions unfold in real life situations and avoid exclusively focusing on observational abilities in lab conditions. However, it is important to notice that the question of how much people actually rely on mindreading in real life social settings is an empirical issue which cannot be settled by philosophical arguments alone. The standard claim that mindreading is ubiquitous in human social cognition, as well as the interactionist claim that people hardly ever mindread, are both speculative. Neither of these claims is supported by empirical evidence. Instead of replacing the study of mindreading with the study of social interactions in the absence of mindreading, I propose to widen the research framework to enable the study of both. Only then can we begin to understand how different components of human social cognition work in ensemble and investigate the specific role of mindreading in human social cognition. The main aim of my dissertation is precisely to lead research of social cognition beyond the current debate between theories of mindreading and interactionism and to outline plausible empirical hypotheses about how different social cognitive abilities function together. A more specific aim of the thesis is to

provide a possible explanation of the function of mindreading in human social cognition.

The first research article ("Toward an integrative account of social cognition: Marrying theory of mind and interactionism to study the interplay of Type 1 and Type 2 processes", see appendix 1) explains why instead of taking the standard theories of mindreading and interactionism to be mutually exclusive opponents, these two approaches should be integrated into a more comprehensive theoretical framework. An integrative theoretical framework is outlined with the help of *dual process models* of social cognition that differentiate between two types of social cognitive processing. The first type (*Type 1*) refers to fast, efficient, stimulus-driven, and relatively inflexible social cognitive processing. The second type (*Type 2*) refers to relatively slow, cognitively laborious, flexible, and possibly partly conscious processing. In the article, it is argued that while interactionism focuses on aspects of social cognition that are likely to be driven by Type 1 processes, theories of mindreading typically target those aspects of social cognition that are likely to be based on Type 2 processes. Furthermore, because the two theoretical frameworks focus on two different aspects of social cognition, they should be seen as complementary rather than mutually exclusive. Based on a critical analysis of recent neuroscientific studies, it is hypothesised that in real life social interactions both types of processes are involved, and thus that human social behaviour may be sustained by the interplay of the two types of process.

The second research article ("Theory of mind and the unobservability of other minds", co-authored with Nivedita Gangopadhyay; see appendix 2) is a conceptual analysis of what it means to talk about the unobservability of other minds. Interactionists frequently claim that the standard theories of mindreading presuppose that mental states are unobservable. They argue that since mental states are embodied and sometimes directly perceivable, theories of mindreading are based on a false assumption. The aim of the article is to show that the interactionist critique is misplaced. In the first part of the article, four different readings of the unobservability claim are outlined: a) a metaphysical reading, concerning what minds are; b) a phenomenological reading, concerning what the phenomenology of mindreading is like; c) an epistemological reading, concerning how to justify beliefs about other minds; and d) a psychological reading, concerning the cognitive mechanisms for mindreading. It is shown that it is *not* the case that theories of mindreading assume that other minds are metaphysically, phenomenologically, or epistemologically unobservable. Theories of mindreading can be said to assume the unobservability of other minds in the psychological sense, i.e. in the sense that they assume that more than perceptual processes are involved in mindreading. The second part of the article lays out a more fine-grained analysis of the psychological reading of the unobservability claim. The psychological reading is cashed out in a strong, medium, and weak version. It is argued that whereas the critics tend to attribute the strongest psychological version of the unobservability claim to theories of mindreading, proponents of theories of mindreading actually subscribe to either the medium

or weak version. It follows that the allegation against theories of mindreading is seriously misdirected. The third part of the article brings out an important constraint on the observability of other minds by using an insight from phenomenology. Based on Husserl's analysis of the structure of perception, it is shown that any theory of social cognition must reckon with the stipulation that mental states cannot be observed in the same way as sensory properties of physical objects. It is shown that theories of mindreading can easily deal with the above-mentioned constraint, and thus that interactionists, far from having proved that theories of mindreading falsely rely on the unobservability assumption, have something to learn from theories of mindreading.

The third research article ("We read minds to shape relationships") looks at the specific *function* of mindreading in human social cognition, and investigates the *motivational mechanisms* that trigger mindreading. Because of the prevalent assumption that mindreading is ubiquitous in human social cognition and serves the general purpose of behaviour explanation and prediction, questions about the function and motivation of mindreading have not arisen as research questions to be studied empirically. In contrast, interactionists claim that mindreading is peripheral to human social cognition, which makes the function and motivation of mindreading appear as not worthwhile to be studied. In article 3, I argue that neither side of the debate captures the specific role of mindreading, which is precisely why this issue needs to be empirically investigated. At the same time, an important aspect of human social cognition has been ignored by both theories of mindreading and interactionism: social relationships. I introduce a theory of the basic structures of social relationships and their cognitive underpinnings – the *relational models theory* of Alan Fiske – and use it as a basis for thinking about the function of explicit mindreading in human social cognition. I put forward the hypothesis that the evolutionary function, as well as the individual purpose, of mindreading is to monitor and shape social relations. More specifically, I hypothesise that people typically mindread 1) when they are uncertain about what kind of social relationship they are dealing with, when interaction becomes problematic, or, is both important and uncertain in outcome; 2) when they wish to change the format of the current relationship or some aspects of it, or apprehend that the other has such a wish; 3) when it is important to anticipate how a current relationship influences and/or is influenced by other relationships; 4) while they make moral judgements (in cultures where making moral judgements entails taking into account whether an act was carried out intentionally or not, as well as other aspects of mental states). A particular method – *experience sampling* – is suggested for testing the hypothesis. In sum, I argue that the function of mindreading is not to predict and explain the behaviour of other people, as if they were another set of objects in the environment; it is to shape social relationships by bringing forth mutual adjustments in the behaviour of oneself and the other party of the relationship. However, saying that mindreading has the function of shaping relationships is not to say that people are motivated to mindread by being aware of its function. Instead, I hypothesise that explicit mindreading may be motivated by social

emotions. Finally, I argue that relational cognition is more fundamental than mindreading in human social cognition.

I would like to think of my dissertation as a snapshot of an ongoing work in progress. I hope that the theoretical considerations I have outlined in my articles will help to lead the research of human social cognition beyond the current debate between theories of mindreading and interactionism, towards an integrative approach that includes insights from both sides of the debate. It is time to move from studying single aspects of social cognition to the study of how different social cognitive abilities work in ensemble, and to open up the research agenda for novel questions, such as "What are the motivating mechanisms that trigger mindreading in social situations?" In particular, I see the study of social relationships and relational cognition as an important part of future research of social cognition. I also hope that the empirical hypotheses presented in the dissertation will inspire empirical scientists to test the claims I have argued for. Hopefully, in the future study of social cognition we will see more collaboration between cognitive psychology, social psychology and anthropology, and I think that philosophers have the potential to initiate the bringing together different disciplines so that they can jointly study different aspects of social cognition.

REFERENCES

Andrews, K. (2009). Understanding norms without a theory of mind. *Inquiry*, *52*(5), 433–448.

Andrews, K. (2012). *Do Apes Read Minds? Toward a New Folk Psychology*. Cambridge, MA: MIT Press.

Apperly, I. (2011). *Mindreaders: The Cognitive Basis of "Theory of Mind."* Hove; New York: Psychology Press.

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.

Astington, J. W., & Gopnik, A. (1991). Developing understanding of desire and intention. In A. Whiten (Ed.), *Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading* (pp. 39–50). Oxford: Basil Blackwell.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*(2), 230–244.

Baron-Cohen, S. (1997). *Mindblindness: An Essay on Autism and Theory of Mind*. Cambridge MA: MIT Press.

Baron-Cohen, S., & Cross, P. (1992). Reading the eyes: Evidence for the role of perception in fhe development of a theory of mind. *Mind & Language*, *7*(1–2), 173–186.

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

Barresi, J., & Moore, C. (1996). Intentional relations and social understanding. *Behavioral and Brain Sciences*, *19*(01), 107–122.

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

Bermúdez, J. L. (2003). The domain of folk psychology. *Royal Institute of Philosophy Supplement*, *53*, 25–48.

Bohl, V. (2011). Milleks on sotsiaalse tunnetuse uurimisvaldkonnas tarvis filosoofiat? *Studia Philosophica Estonica*, *4*(1), 20–51.

Bohl, V. (2014). We read minds to shape relationships. *Philosophical Psychology*, 1–21. Published online: 12 March 2014.

Bohl, V., & Gangopadhyay, N. (2013). Theory of mind and the unobservability of other minds. *Philosophical Explorations*, 1–20. Published online: 30 July 2013.

Bohl, V., & van den Bos, W. (2012). Toward an integrative account of social cognition: Marrying theory of mind and interactionism to study the interplay of Type 1 and Type 2 processes. *Frontiers in Human Neuroscience*, *6*, 1–15. Published online: 11 October 2012.

Brook, A. (2009). Introduction: Philosophy in and philosophy of cognitive science. *Topics in Cognitive Science*, *1*(2), 216–230.

Buber, M. (1923). *Ich und Du*. Leipzig: Insel Verlag.

Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition*, *112*(2), 337–342.

Carruthers, P. (1996). Simulation and self-knowledge: A defence of theory-theory. In P. Carruthers & P. K. Smith (Eds.), *Theories of Theories of Mind* (pp. 22–38). Cambridge: Cambridge University Press.

Carruthers, P. (2006). *The Architecture of the Mind: Massive Modularity and the Flexibility of Thought*. Oxford; New York: Clarendon Press; Oxford University Press.
Cavell, S. (1976). *Must We Mean What We Say?* Cambridge: Cambridge University Press.
Chaminade, T., & Decety, J. (2002). Leader or follower? Involvement of the inferior parietal lobule in agency. *Neuroreport*, *13*(15), 1975–1978.
Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *The Journal of Philosophy*, *78*(2), 67–90.
Churchland, P. M. (1984). *Matter and Consciousness: A Contemporary Introduction to the Philosophy of Mind*. Cambridge, MA: MIT Press.
Cleret de Langavant, L., Jacquemot, C., Bachoud-Lévi, A.-C., & Dupoux, E. (2013). The second person in "I"-"you"-"it" triadic interactions. *Behavioral and Brain Sciences*, *36*(04), 416–417.
Cole, J. (2010). Agency with impairments of movement. In D. Schmicking & S. Gallagher (Eds.), *Handbook of Phenomenology and Cognitive Science* (pp. 655–670). Dordrecht: Springer.
Cruz, J., & Gordon, R. M. (2002). Simulation theory. In (L. Nadel, Ed.) *Encyclopedia of Cognitive Science.* Basingstoke: Palgrave Macmillan.
Currie, G., & Sterelny, K. (2000). How to think about the modularity of mind-reading. *The Philosophical Quarterly*, *50*(199), 145–160.
De Bruin, L. C., & Strijbos, D. W. (2010). Folk psychology without principles: An alternative to the belief–desire model of action interpretation. *Philosophical Explorations*, 13(3), 257–274.
De Jaegher, H. (2009b). Social understanding through direct perception? Yes, by interacting. *Consciousness and Cognition*, *18*(2), 535–542.
De Jaegher, H. (2009a). What made me want the cheese? A reply to Shaun Gallagher and Dan Hutto. *Consciousness and Cognition*, *18*(2), 549–550.
De Jaegher, H., & Di Paolo, E. (2013). Enactivism is not interactionism. *Frontiers in Human Neuroscience*, 6, 1–2. Published online: 3 January 2013.
De Jaegher, H., & Di Paolo, E. (2007). Participatory sense-making: An enactive approach to social cognition. *Phenomenology and the Cognitive Sciences*, 485–507.
De Jaegher, H., Di Paolo, E., & Gallagher, S. (2010). Can social interaction constitute social cognition? *Trends in Cognitive Sciences*, *14*(10), 441–447.
De Vignemont, F. (2009). Drawing the boundary between low-level and high-level mindreading. *Philosophical Studies*, *144*(3), 457–466.
Dennett, D. C. (1969). *Content and Consciousness*. London; New York: Routledge.
Dennett, D. C. (1978). Beliefs about beliefs. *Behavioral and Brain Sciences*, *1*(4), 568.
Dennett, D. C. (1987). *The Intentional Stance*. Cambridge, MA: MIT Press.
Dennett, D. C. (2009). The part of cognitive science that is philosophy. *Topics in Cognitive Science*, *1*(2), 231–236.
Descartes, R. (1998). *Discourse on Method and Meditations on First Philosophy* (4th ed.). Indianapolis: Hackett Pub.
Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: A neurophysiological study. *Experimental Brain Research*, *91*(1), 176–180.
Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE*, *7*(1), e29081.
Doyle, A. C. (1894). The Musgrave ritual. In *The Memoirs of Sherlock Holmes*. New York: Harper Bros.

Dullstein, M. (2012). The second person in the theory of mind debate. *Review of Philosophy and Psychology*, *3*(2), 231–248.
Evans, N. (2013). On projecting grammatical persons into social neurocognition: A view from linguistics. *Behavioral and Brain Sciences*, *36*(04), 419–420.
Fadiga, L., Fogassi, L., Pavesi, G., & Rizzolatti, G. (n.d.). Motor facilitation during action observation: A magnetic stimulation study. *Journal of Neurophysiology*, *73*(6), 2608–2611.
Farrer, C., & Frith, C. D. (2002). Experiencing oneself vs another person as being the cause of an action: The neural correlates of the experience of agency. *NeuroImage*, *15*(3), 596–603.
Flavell, J. H. (1988). The development of children's knowledge about the mind: From cognitive connections to mental representations. In J. W. Astington, P. L. Harris, & D. R. Olson (Eds.), *Developing Theories of Mind* (pp. 244–267). New York: Cambridge University Press.
Fodor, J. A. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MA: MIT Press.
Fodor, J. A. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press.
Frith, U., & Happe, F. (1999). Theory of mind and self-consciousness: What is it like to be autistic? *Mind & Language*, *14*(1), 82–89.
Froese, T., & Gallagher, S. (2012). Getting interaction theory (IT) together: Integrating developmental, phenomenological, enactive, and dynamical approaches to social interaction. *Interaction Studies*, *13*(3), 436–468.
Gallagher, S. (2001). The practice of mind: Theory, simulation or primary interaction? *Journal of Consciousness Studies*, *8*(5–7), 83–108.
Gallagher, S. (2003). Phenomenology and experimental design. *Journal of Consciousness Studies*, *10*(9–10), 85–99.
Gallagher, S. (2004). Understanding interpersonal problems in autism: Interaction theory as an alternative to theory of mind. *Philosophy, Psychiatry, & Psychology*, *11*(3), 199–217.
Gallagher, S. (2005). Phenomenological contributions to a theory of social cognition. *Husserl Studies*, *21*(2), 95–110.
Gallagher, S. (2007). Simulation trouble. *Social Neuroscience*, *2*(3–4), 353–365.
Gallagher, S. (2008a). Another look at intentions: A response to Raphael van Riel's "On how we perceive the social world." *Consciousness and Cognition*, *17*(2), 553–555.
Gallagher, S. (2008b). Direct perception in the intersubjective context. *Consciousness and Cognition*, *17*(2), 535–543.
Gallagher, S. (2008c). Inference or interaction: Social cognition without precursors. *Philosophical Explorations*, *11*(3), 163–174.
Gallagher, S. (2008d). Intersubjectivity in perception. *Continental Philosophy Review*, *41*(2), 163–178.
Gallagher, S. (2008e). Understanding others: Embodied social cognition. In P. Calvo & A. Gomila (Eds.), *Handbook of Cognitive Science: An Embodied Approach*. Amsterdam: Elsevier.
Gallagher, S. (2009). Two problems of intersubjectivity. *Journal of Consciousness Studies*, *16*(6–8), 289–308.
Gallagher, S. (2012). In defense of phenomenological approaches to social cognition: Interacting with the critics. *Review of Philosophy and Psychology*, *3*(2), 187–212.
Gallagher, S., & Hutto, D. D. (2008). Understanding others through primary interaction and narrative practice. In J. Zlatev, T. Racine, C. Sinha, & E. Itkonen (Eds.), *The*

Shared Mind: Perspectives on Intersubjectivity (pp. 17–38). Amsterdam: John Benjamins Publishing Company.
Gallagher, S., & Meltzoff, A. N. (1996). The earliest sense of self and others: Merleau-Ponty and recent developmental studies. *Philosophical Psychology*, *9*(2), 211–233.
Gallagher, S., & Zahavi, D. (2008). *The Phenomenological Mind*. London; New York: Routledge.
Gallese, V., & Goldman, A. I. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, *2*(12), 493–501.
Gerrans, P. (2002). The theory of mind module in evolutionary psychology. *Biology and Philosophy*, *17*(3), 305–321.
Gibson, J. J. (1977). The theory of affordances. In R. Shaw & J. Bransford (Eds.), *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*. Hillsdale: Erlbaum.
Gibson, J. J. (1979). *The Ecological Approach to Visual Perception.* Boston: Houghton Mifflin.
Godfrey-Smith, P. (2005). Folk psychology as a model. *Philosopher's Imprint*, *5*(6), 1–16.
Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. *Trends in Cognitive Sciences*, *3*(11), 419–429.
Goldman, A. I. (1989). Interpretation psychologized. *Mind & Language*, *4*(3), 161–185.
Goldman, A. I. (1992). In defence of the simulation theory. *Mind & Language*, *7*(1–2), 104–119.
Goldman, A. I. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford; New York: Oxford University Press.
Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences*, *16*(01), 1–14.
Gopnik, A. (1996). Theories and modules: Creation myths, developmental realities, and Neurath's boat. In P. Carruthers & P. K. Smith (Eds.), *Theories of Theories of Mind* (pp. 169–183). New York: Cambridge University Press.
Gopnik, A. (1998). Explanation as orgasm. *Minds and Machines*, *8*(1), 101–118.
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.
Gopnik, A., & Meltzoff, A. N. (1997). *Words, Thoughts, and Theories*. Cambridge, MA: MIT Press.
Gopnik, A., Meltzoff, A. N., & Kuhl, P. K. (2001). *The Scientist in the Crib: Minds, Brains, and How Children Learn*. New York: Harper Perennial.
Gopnik, A., Slaughter, V., & Meltzoff, A. N. (1994). Toddlers' understanding of intentions, desires, and emotions: Explorations of the dark ages. In C. Lewis & P. Mitchell (Eds.), *Children's Early Understanding of Mind: Origins and Development* (pp. 157–181). Hillsdale: Lawrence Erlbaum Associates.
Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind & Language*, *7*(1–2), 145–171.
Gordon, R. M. (1986). Folk psychology as simulation. *Mind & Language*, *1*(2), 158–171.
Gordon, R. M. (1995). Simulation without introspection or inference from me to you. In T. Stone & M. Davies (Eds.), *Mental Simulation: Evaluations and Applications* (pp. 53–67). Oxford, UK; Cambridge, MA: Blackwell.
Gordon, R. M. (1996). Radical simulation. In P. Carruthers & P. K. Smith (Eds.), *Theories of Theories of Mind* (pp. 11–21). Cambridge: Cambridge University Press.

Harman, G. (1978). Studying the chimpanzee's theory of mind. *Behavioral and Brain Sciences*, *1*(4), 576.
Harris, P. L. (1992). From simulation to folk psychology: The case for development. *Mind & Language*, *7*(1–2), 120–144.
Heal, J. (1986). Replication and functionalism. In J. Butterfield (Ed.), *Language, Mind, and Logic* (pp. 135–150). Cambridge: Cambridge University Press.
Heal, J. (1995). How to think about thinking. In T. Stone & M. Davies (Eds.), *Mental Simulation: Evaluations and Applications* (pp. 33–52). Oxford, UK; Cambridge, MA: Blackwell.
Heal, J. (1996). Simulation, theory, and content. In P. Carruthers & P. K. Smith (Eds.), *Theories of Theories of Mind* (pp. 75–89). Cambridge: Cambridge University Press.
Heal, J. (1998). Co-Cognition and off-line simulation: Two ways of understanding the simulation approach. *Mind & Language*, *13*(4), 477–498.
Heal, J. (2003). *Mind, Reason, and Imagination: Selected Essays in Philosophy of Mind & Language*. Cambridge, UK; New York: Cambridge University Press.
Heal, J. (2005). Review of *Mindreading: An Integrated Account of Pretence, Self-Awareness and Understanding Other Minds*, by Shaun Nichols and Stephen P. Stich. Oxford: Clarendon Press, 2003. Pp. 237. *Mind*, *114*, 181–184.
Herschbach, M. (2008). Folk psychological and phenomenological accounts of social perception. *Philosophical Explorations*, *11*(3), 223–235.
Herschbach, M. (2012). On the role of social interaction in social cognition: A mechanistic alternative to enactivism. *Phenomenology and the Cognitive Sciences*, *11*(4), 467–486.
Hutto, D. D. (2004). The limits of spectatorial folk psychology. *Mind & Language*, *19*(5), 548–573.
Hutto, D. D. (2008a). *Folk Psychological Narratives: The Sociocultural Basis of Understanding Reasons*. Cambridge, MA: MIT Press.
Hutto, D. D. (2008b). The Narrative Practice Hypothesis: Clarifications and implications. *Philosophical Explorations*, *11*(3), 175–192.
Hutto, D. D. (2009). ToM rules, but it is not OK! In I. Leudar & A. Costall (Eds.), *Against Theory of Mind* (pp. 221–238). Basingstoke: Palgrave Macmillan.
Iacoboni, M., & Mazziotta, J. C. (2007). Mirror neuron system: Basic findings and clinical applications. *Annals of Neurology*, *62*(3), 213–218.
Knobe, J. (2008). The concept of intentional action: A case study in the uses of folk psychology. In J. M. Knobe & S. Nichols (Eds.), *Experimental Philosophy* (pp. 129–148). Oxford: Oxford University Press.
Knobe, J., & Nichols, S. (2007). An experimental philosophy manifesto. In *Experimental Philosophy* (pp. 3–16). Oxford: Oxford University Press.
Knudsen, B., & Liszkowski, U. (2012). 18-month-olds predict specific action mistakes through attribution of false belief, not ignorance, and intervene accordingly. *Infancy*, *17*(6), 672–691.
Krueger, J. (2013). Merleau-Ponty on shared emotions and the joint ownership thesis. *Continental Philosophy Review*, *46*(4), 509–531.
Krueger, J., & Overgaard, S. (2012). Seeing subjectivity: Defending a perceptual account of other minds. *ProtoSociology: Consciousness and Subjectivity*, *47*, 239–262.
Leslie, A. M. (1987). Pretense and representation: The origins of "theory of mind." *Psychological Review*, *94*(4), 412–426.

Leslie, A. M. (1994). ToMM, ToBY, and Agency: Core architecture and domain specificity. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the Mind* (pp. 119–148). Cambridge: Cambridge University Press.

Leslie, A. M., & Roth, D. (1993). What autism teaches us about metarepresentation. In S. Baron-Cohen, H. T. Flusberg, & H. Cohen (Eds.), *Understanding Other Minds: Perspectives from Autism* (pp. 83–111). Oxford: Oxford University Press.

Leudar, I., & Costall, A. (Eds.). (2009a). *Against Theory of Mind*. Basingstoke; New York: Palgrave Macmillan.

Leudar, I., & Costall, A. (2009b). Introduction: Against "Theory of Mind." In I. Leudar & A. Costall (Eds.), *Against Theory of Mind* (pp. 1–16). Basingstoke; New York: Palgrave Macmillan.

Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy*, *50*(3), 249–258.

Lillard, A. (1998). Ethnopsychologies: Cultural variations in theories of mind. *Psychological Bulletin*, *123*(1), 3–32.

Low, J., & Perner, J. (2012). Implicit and explicit theory of mind: State of the art: Editorial. *British Journal of Developmental Psychology*, *30*(1), 1–13.

Lutz, A. (2002). Toward a neurophenomenology as an account of generative passages: A first empirical case study. *Phenomenology and the Cognitive Sciences*, *1*(2), 133–2002.

Lutz, A., Lachaux, J.-P., Martinerie, J., & Varela, F. (2001). Guiding the study of brain dynamics by using first-person data: Synchrony patterns correlate with ongoing conscious states during a simple visual task. *Proceedings of the National Academy of Sciences of the USA*, *99*(2), 1586–1591.

Maibom, H. (2003). The mindreader and the scientist. *Mind & Language*, *18*(3), 296–315.

Maibom, H. (2007). Social systems. *Philosophical Psychology*, *20*(5), 557–578.

Mameli, M. (2001). Mindreading, mindshaping, and evolution. *Biology and Philosophy*, *16*(5), 595–626.

McGeer, V. (2007). The regulative dimension of folk psychology. In M. Ratcliffe & D. D. Hutto (Eds.), *Folk Psychology Re-Assessed* (pp. 137–156). Dordrecht; London: Springer.

Meltzoff, A. N., & Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science*, *198*(4312), 75–78.

Meltzoff, A. N., & Moore, M. K. (1983). Newborn infants imitate adult facial gestures. *Child Development*, *54*(3), 702–709.

Meltzoff, A. N., & Moore, M. K. (1989). Imitation in newborn infants: Exploring the range of gestures imitated and the underlying mechanisms. *Developmental Psychology*, *25*(6), 954–962.

Merleau-Ponty, M. (2002). *Phenomenology of Perception*. London; New York: Routledge.

Michael, J. (2011). Interactionism and mindreading. *Review of Philosophy and Psychology*, *2*(3), 559–578.

Morton, A. (1980). *Frames of Mind: Constraints on the Common-Sense Conception of the Mental*. Oxford; New York: Clarendon Press; Oxford University Press.

Morton, A. (1996). Folk psychology is not a predictive device. *Mind*, *105*(417), 119–137.

Morton, A. (2003). *The Importance of Being Understood: Folk Psychology as Ethics*. London; New York: Routledge.

Morton, A. (2009). Folk psychology. In A. Beckermann, B. P. McLaughlin, & S. Walter (Eds.), *The Oxford Handbook of Philosophy of Mind*. Oxford University Press.

Mukamel, R., Ekstrom, A. D., Kaplan, J., Iacoboni, M., & Fried, I. (2010). Single-neuron responses in humans during execution and observation of actions. *Current Biology*, *20*(8), 750–756.

Murray, L., & Trevarthen, C. (1985). Emotional regulation of interactions between two-month-olds and their mothers. In T. M. Field & N. A. Fox (Eds.), *Social Perception in Infants* (pp. 177–197). Norwood, NJ: Ablex.

Nadel, J., Carchon, I., Kervella, C., Marcelli, D., & Reserbat-Plantey, D. (1999). Expectancies for social contingency in 2-month-olds. *Developmental Science*, *2*(2), 164–173.

Newen, A., & Schlicht, T. (2009). Understanding other minds: A criticism of Goldman's simulation theory and an outline of the person model theory. *Grazer Philosophische Studien*, *79*, 209–242.

Nichols, S., & Stich, S. P. (2003). *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*. Oxford; New York: Oxford University Press.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*(5719), 255–258.

Overgaard, S., & Michael, J. (2013). The interactive turn in social cognition research: A critique. *Philosophical Psychology*, 1–24.

Papineau, D. (2009a). Naturalism. In (E. N. Zalta, Ed.) *The Stanford Encyclopedia of Philosophy (Spring 2009 Edition)*.

Papineau, D. (2009b). The poverty of analysis. *Aristotelian Society Supplementary Volume*, *83*(1), 1–30.

Perner, J. (1988). Developing semantics for theories of mind: From propositional attitudes to mental representations. In J. W. Astington, P. L. Harris, & D. R. Olson (Eds.), *Developing Theories of Mind* (pp. 141–172). New York: Cambridge University Press.

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

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(04), 515.

Prinz, J. (2008). Empirical philosophy and experimental philosophy. In *Experimental Philosophy*. Oxford: Oxford University Press.

Pylyshyn, Z. W. (1978). When is attribution of beliefs justified? *Behavioral and Brain Sciences*, *1*(4), 592.

Ratcliffe, M. (2006). 'Folk psychology' is not folk psychology. *Phenomenology and the Cognitive Sciences*, *5*(1), 31–52.

Ratcliffe, M. (2007). *Rethinking Commonsense Psychology: A Critique of Folk Psychology, Theory of Mind and Simulation*. Basingstoke: Palgrave Macmillan.

Ravenscroft, I. (2010). Folk psychology as a theory. In (E. N. Zalta, Ed.) *The Stanford Encyclopedia of Philosophy (Fall 2010 Edition)*.

Reddy, V. (1996). Omitting the second person in social understanding. *Behavioral and Brain Sciences*, *19*(01), 140–141.

Reddy, V. (2008). *How Infants Know Minds*. Cambridge, MA; London: Harvard University Press.

Reddy, V., & Morris, P. (2004). Participants don't need theories: Knowing minds in engagement. *Theory and Psychology*, *14*(5), 647–665.

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: Evidence from 14- and 18-moth-olds. *Developmental Psychology*, *33*(1), 12–21.
Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, *3*, 131–141.
Robbins, P. (2010). Modularity of Mind. In (E. N. Zalta, Ed.) *The Stanford Encyclopedia of Philosophy (Summer 2010 Edition)*.
Samuels, R. (1998). Evolutionary psychology and the massive modularity hypothesis. *The British Journal for the Philosophy of Science*, *49*(4), 575–602.
Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience. *Behavioral and Brain Sciences*, *36*(04), 393–414.
Schilbach, L., Wohlschlaeger, A. M., Kraemer, N. C., Newen, A., Shah, N. J., Fink, G. R., & Vogeley, K. (2006). Being with virtual others: Neural correlates of social interaction. *Neuropsychologia*, *44*(5), 718–730.
Scholl, B. J., & Leslie, A. M. (1999). Modularity, development, and "Theory of Mind." *Mind & Language*, *14*(1), 131–153.
Segal, G. (1996). The modularity of theory of mind. In P. Carruthers & P. K. Smith (Eds.), *Theories of Theories of Mind* (pp. 141–157). Cambridge: Cambridge University Press.
Sellars, W. (1962). Philosophy and the scientific image of man. In R. G. Colodny (Ed.), *Frontiers of Science and Philosophy* (pp. 35–78). Pittsburgh: University of Pittsburgh Press.
Sellars, W. (1997). *Empiricism and the Philosophy of Mind*. Cambridge, MA: Harvard University Press.
Shanahan, M. (2009). The frame problem. In (E. N. Zalta, Ed.) *The Stanford Encyclopedia of Philosophy (Winter 2009 Edition)*.
Sharrock, W., & Coulter, J. (2004). ToM: A critical commentary. *Theory and Psychology*, *14*(5), 579–600.
Slors, M. (2010). Neural resonance: Between implicit simulation and social perception. *Phenomenology and the Cognitive Sciences*, *9*(3), 437–458.
Slors, M. (2012). The model-model of the theory-theory. *Inquiry*, *55*(5), 521–542.
Spaulding, S. (2010). Embodied cognition and mindreading. *Mind & Language*, *25*(1), 119–140.
Spaulding, S. (2012). Introduction to debates on embodied social cognition. *Phenomenology and the Cognitive Sciences*, *11*(4), 431–448.
Stich, S., & Nichols, S. (1992). Folk psychology: Simulation or tacit theory? *Mind & Language*, *7*(1), 35–71.
Stich, S., & Nichols, S. (1997). Cognitive penetrability, rationality and restricted simulation. *Mind & Language*, *12*(3–4), 297–326.
Stich, S., & Nichols, S. (1998). Theory theory to the max. *Mind & Language*, *13*(3), 421–449.
Stich, S., & Nichols, S. (2007). Cognitive penetrability, rationality and restricted simulation. *Mind & Language*, *12*(3–4), 297–326.
Stueber, K. R. (2006). *Rediscovering Empathy: Agency, Folk Psychology, and the Human Sciences*. Cambridge, MA: MIT Press.
Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science*, *18*(7), 580–586.
Tager-Flusberg, H. (2005). What neurodevelopmental disorders can reveal about cognitive architecture: The example of theory of mind. In P. Carruthers, S. Laurence,

& S. Stich (Eds.), *The Innate Mind: Structure and Contents* (pp. 272–288). New York: Oxford University Press.
Thagard, P. (2009). Why cognitive science needs philosophy and vice versa. *Topics in Cognitive Science*, *1*(2), 237–254.
Tooming, U. (2013). Without pretense: A critique of Goldman's model of simulation. *Phenomenology and the Cognitive Sciences*.
Trevarthen, C. (1979). Communication and cooperation in early infancy. A description of primary intersubjectivity. In M. Bullowa (Ed.), *Before Speech: The Beginning of Human Communication* (pp. 321–347). Cambridge: Cambridge University Press.
Trevarthen, C., & Hubley, P. (1978). Secondary intersubjectivity: Confidence, confiding and acts of meaning in the first year. In A. Lock (Ed.), *Action, Gesture and Symbol: The Emergence of Language,* (pp. 183–229). London: Academic Press.
Van Gelder, T. (1998). The roles of philosophy in cognitive science. *Philosophical Psychology*, *11*(2), 117–136.
Varela, F. (1996). Neurophenomenology: A methodological remedy to the hard problem. *Journal of Consciousness Studies*, *3*(4), 330–350.
Vignemont, F. (2009). Drawing the boundary between low-level and high-level mindreading. *Philosophical Studies*, *144*(3), 457–466.
Weinberg, M. K., & Tronick, E. Z. (1996). Infant affective reactions to the resumption of maternal interaction after the still-face. *Child Development*, *67*(3), 905–914.
Wellman, H. M. (1990). *The Child's Theory of Mind*. Cambridge, MA: MIT Press.
Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, *72*(3), 655–684.
Wilms, M., Schilbach, L., Pfeiffer, U., Bente, G., Fink, G. R., & Vogeley, K. (2010). It's in your eyes–using gaze-contingent stimuli to create truly interactive paradigms for social cognitive and affective neuroscience. *Social Cognitive and Affective Neuroscience*, *5*(1), 98–107.
Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128.
Yirmiya, N., Solomonica-Levi, D., Shulman, C., & Pilowsky, T. (1996). Theory of mind abilities in individuals with autism, Down syndrome, and mental retardation of unknown etiology: The role of age and intelligence. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *37*(8), 1003–1014.
Zahavi, D. (2004). Phenomenology and the project of naturalization. *Phenomenology and the Cognitive Sciences*, *3*(4), 331–347.
Zahavi, D. (2005). *Subjectivity and Selfhood: Investigating the First-Person Perspective*. Cambridge, MA: MIT Press.
Zahavi, D. (2007). Expression and empathy. In D. D. Hutto & M. Ratcliffe (Eds.), *Folk Psychology Re-Assessed* (pp. 25–40). Dordrecht: Springer Netherlands.
Zahavi, D. (2008). Simulation, projection and empathy. *Consciousness and Cognition*, *17*(2), 514–522.
Zahavi, D. (2010). Naturalized phenomenology. In S. Gallagher & D. Schmicking (Eds.), *Handbook of Phenomenology and Cognitive Science* (pp. 2–19). Berlin: Springer.
Zahavi, D. (2011). Empathy and direct social perception: A phenomenological proposal. *Review of Philosophy and Psychology*, *2*(3), 541–558.
Zahavi, D., & Gallagher, S. (2008). The (in)visibility of others: A reply to Herschbach. *Philosophical Explorations*, *11*(3), 237–244.

Zahavi, D., & Parnas, J. (2003). Conceptual problems in infantile autism research: Why cognitive science needs phenomenology. *Journal of Consciousness Studies*, *9*(10), 53–71.

Zawidzki, T. W. (2008). The function of folk psychology: Mind reading or mind shaping? *Philosophical Explorations*, *11*(3), 193–210.

Zawidzki, T. (2013). *Mindshaping: A New Framework for Understanding Human Social Cognition*. Cambridge, MA: MIT Press.

Zlatev, J., Racine, T., Sinha, C., & Itkonen, E. (2008f). Intersubjectivity: What makes us human? In J. Zlatev, T. Racine, C. Sinha, & E. Itkonen (Eds.), *The Shared Mind: Perspectives on Intersubjectivity* (pp. 1–14). Amsterdam; Philadelphia: John Benjamins Publishing Company.

PUBLICATIONS

SUMMARY IN ESTONIAN

Kuidas me teisi mõistame?
Teispool vaimulugemise teooriaid ja interaktsionismi

Käesolev väitekiri uurib inimese sotsiaalset tunnetust ehk küsimust, millised tunnetusprotsessid võimaldavad inimestevahelist interaktsiooni ja mõistmist. Väitekirja peamiseks eesmärgiks on näidata, et seni teineteist välistavateks peetud lähenemisi – *vaimulugemise teooriaid* ja *interaktsionismi* – tuleks käsitleda teineteist täiendavatena. Esitatakse neid kahte lähenemist ühendav ja ületav teoreetiline raamistik, mis tõstatab uusi olulisi küsimusi ja annab aluse uutele empiirilistele hüpoteesidele sotsiaalsete tunnetusprotsesside kohta.

Väitekiri koosneb pikemast ülevaateartiklist ja kolmest eelretsenseeritud artiklist, mis on ilmunud rahvusvahelistes teadusajakirjades. Ülevaateartikkel algab metodoloogilise peatükiga, kus käsitletakse erinevaid filosoofilisi meetodeid, mida saab rakendada sotsiaalse tunnetuse interdistsiplinaarsel uurimisel. Peatükis arutlen filosoofia rolli üle sotsiaalse tunnetuse uurimisel üldiselt ja selgitan, kuidas olen filosoofilisi meetodeid rakendanud väitekirja kuuluvates teadusartiklites. Pean filosoofia-alast pädevust oluliseks, sest see võimaldab 1) teaduslike paradigmade aluseks olevaid teoreetilisi ja metodoloogilisi eeldusi kriitiliselt analüüsida; 2) näha eri teadusharude lõikumispunktides uusi küsimusi, ergutades seeläbi uute empiiriliste hüpoteeside ja uurimismeetodite väljatöötamist; 3) reflekteerida selle üle, kuidas erinevate teadusharude ja paradigmade empiirilised uurimistulemused ja teoreetilised väited omavahel kokku sobivad, viies seeläbi uute laiahaardelisemate teoreetiliste raamistikeni; 4) korrastada mõistevõrgustikke ja töötada välja uusi teoreetilisi mõisteid. Eraldi alapeatükk on pühendatud fenomenoloogia rollile sotsiaalse tunnetuse uurimisel. Toon välja, et fenomenoloogiat saab rakendada vähemalt neljal moel: 1) tõlgendades olemasolevaid empiirilisi andmeid fenomenoloogiast lähtuvalt; 2) kasutades fenomenoloogilisi kirjeldusi *eksplanandumina* sotsiaaltunnetuslike protsesside uurimisel; 3) neurofenomenoloogiliste eksperimentide läbiviimisel, kus subjektide fenomenoloogiliste kirjelduste põhjal loodud kategooriaid kasutatakse andmete kategoriseerimisel; 4) fenomenoloogiliste väidete ”eellaadimisel” eksperimentidesse.

Ülevaateartikli teine peatükk algab olulisemate vaimulugemise eksperimentide (nt vääruskumuse testide) tutvustamisega. Seejärel annan ülevaate peavoolu teooriatest vaimuseisundite (uskumuste, soovide, kavatsuste, emotsioonide jne) omistamise aluseks olevate tunnetuslike protsesside kohta: *teooria-teooria* (Alison Gopnik, Andrew Meltzoff, Henry M. Wellmann, Josef Perner, John H. Flavell, Alan Leslie, Simon Baron-Cohen), *simulatsiooniteooria*, (Jane Heal, Robert Gordon, Alvin Goldman) ja neid kahte ühendavate *hübriidteooriate* (Shaun Nichols, Stephen Stich, Alvin Goldman) tuntumatest versioonidest. Teooria-teooria järgi toetuvad inimesed vaimuseisundite omistamisel rahvapsühholoogilisele teooriale selle kohta, kuidas eri tüüpi vaimuseisundid on seotud üksteisega ja käitumisega. Teooria-teoorial on laias laastus kaks

versiooni: *lapsteadlase teooriad* ja *modulaarsed teooriad*. Lapsteadlase teooriate kohaselt omandavad lapsed rahvapsühholoogilise teooria analoogselt sellega, kuidas teadlased konstrueerivad teaduslikke teooriaid. Modulaarsete teooriate järgi seisneb rahvapsühholoogilise teooria omandamine vastava kaasasündinud kognitiivse mooduli küpsemises. Simulatsiooniteooria seevastu ütleb, et teistele inimestele vaimuseisundite omistamisel ei lähtuta teooriast, vaid kasutatakse omaenda vaimset masinavärki mudelina teiste inimeste vaimuseisundite simuleerimiseks: kujutletakse end teise inimese olukorda ja omistatakse selle tulemusel esile kerkinud vaimuseisundid inimesele, keda soovitakse mõista. Hiljuti on tekkinud ka hübriidteooriad, mis kombineerivad simulatsiooniteooria ja teooria-teooria elemente.

Ülevaateartikli kolmas peatükk tutvustab viimasel kümnendil esile kerkinud vaimulugemise teooriate suhtes kriitilist lähenemist: *interaktsionismi* (Shaun Gallagher, Vasudevi Reddy, Hanne De Jaegher, Ezequiel Di Paolo, Matthew Ratcliffe, Daniel D. Hutto, Ivan Leudar, Alan Costall, Dan Zahavi). Interaktsionistid väidavad, et sotsiaalse tunnetuse uurimisel tuleb keskenduda vaimulugemise uurimise asemel sotsiaalse interaktsiooni uurimisele. Interaktsionistid lükkavad tagasi vaimulugemise teoreetikute eelduse, et vaimuseisundite omistamine ehk vaimulugemine on võtmeteguriks, mis võimaldab inimestel üksteist mõista ja sotsiaalselt interakteeruda. Nad väidavad, et inimestevaheline suhtlus toimub enamasti ilma vaimulugemiseta, ning et sotsiaalse tunnetuse aluseks on tajulised, sensorimotoorsed ja afektiivsed protsessid, millele arengu käigus lisandub jagatud sotsiaalse konteksti mõistmise võime. Vaimulugemisele jääb sotsiaalses tunnetuses üksnes marginaalne roll: interaktsionistid väidavad, et inimesed omistavad vaimuseisundeid vaid siis, kui teine isik käitub kummaliselt ja muud meetodid tema mõistmiseks ei ole vilja kandnud. Interaktsionistid väidavad sageli, et vaimulugemise teooriad põhinevad eeldusel, et teiste olendite vaim on vaadeldamatu. Kuivõrd nende järgi on vaimuseisundid lahutamatult seotud nende ihulise väljendusega, siis osutub suur osa vaimuseisunditest vähemalt osaliselt vaadeldavaks, millest tulenevalt väidetakse, et vaimulugemise teooriad põhinevad vääral eeldusel. Lisaks rõhutavad interaktsionistid asjaolu, et vaimulugemise teooriad tuginevad eksperimentidele, kus uurimissubjektid pelgalt vaatlevad sotsiaalseid olukordi kõrvalt, kolmanda isiku vaatepunktist, kuid seejuures need teooriad ignoreerivad sotsiaalset tunnetust teise isiku vaatpunktist, mis on olemuslik vahetule sotsiaalsele interaktsioonile. Seetõttu on ebaselge, kas vaimulugemise teooriate aluseks olevad empiirilised uurimused on üldse pädevad sotsiaalse interaktsiooni seletamiseks. Interaktsionistide väitel ei erine sotsiaalses interaktsioonis osalemise ja pelga pealtvaatamise puhul mitte üksnes individuaalsed tunnetusprotsessid, vaid esimesel juhul mängib interaktsiooniprotsess ka sotsiaalset tunnetust võimaldavat ja konstitueerivat rolli, mida tal teisel juhul olla ei saa.

Ma nõustun suuresti interaktsionistliku kriitikaga selles osas, et vaimulugemise teooriate seletusvõime on oluliselt piiratud, ning nad ei ammenda kaugeltki inimese sotsiaalse tunnetuse aluseks olevate protsesside kogu spektrit. Lisaks laboritingimustes läbiviidavatele eksperimentidele, mis keskenduvad

vaimuseisundite omistamisele kolmanda isiku perspektiivist, tuleks kindlasti uurida sotsiaalset suhtlust selle erinevates loomulikes kontekstides. Samas on oluline tähele panna, et küsimus vaimulugemise rollist inimese sotsiaalses tunnetuses on empiiriline küsimus, millele ei saa vastata üksnes filosoofiliste argumentide põhjal. Interaktsionistide väide, et vaimulugemine on sotsiaalse tunnetuse suhtes perifeerse tähtsusega, on sama spekulatiivne kui levinud aruaam, et vaimulugemisel on sotsiaalses tunnetuses läbiv ja keskne roll. Kummagi väite toetuseks ei leidu hetkel piisavalt empiirilisi tõendeid. Ma ei poolda interaktsionistide ettepanekut keskenduda vaimulugemise uurimise asemel sotsiaalse interaktsiooni uurimisele. Palju viljakam oleks ehitada neist kahest teineteisele vastanduvast lähenemisest laiahaardelisem teoreetiline raamistik, mis laseks uurida nii vaimuseisundite omistamist kui sotsiaalse interaktsiooni aluseks olevaid mittementalistlikke sotsiaaltunnetuslikke võimeid. See võimaldaks uurida, kuidas sotsiaalse tunnetuse erinevad komponendid üheskoos funktsioneerivad ning vastata küsimusele, mis on vaimulugemise roll sotsiaalses tunnetuses. Käesoleva väitekirja peamiseks eesmärgiks ongi eeltoodud kahe lähenemise integreerimine. Väitekirja kitsam eesmärk on pakkuda välja empiiriliselt kontrollitav hüpotees eksplitsiitse vaimulugemise funktsioonist inimese sotsiaalses tunnetuses.

Artikkel 1. Esimene artikkel – "Toward an integrative account of social cognition: Marrying theory of mind and interactionism to study the interplay of Type 1 and Type 2 processes" (vt lisa 1) – on kirjutatud koostöös neuroteadlase Wouter van den Bosiga, kes töötab praegu Max Plancki inimarengu instituudis Berliinis. Artikkel ilmus ajakirja *Frontiers in Human Neuroscience* erinumbris, mis keskendus sotsiaalse interaktsiooni neuroteaduslikule uurimisele. Artiklis selgitatakse, miks ei tuleks vaimulugemise teooriaid ja interaktsionismi käsitleda teineteist välistavatena. Esitatakse neid kaht lähenemist ühendav teoreetiline raamistik duaalsete protsesside mudelite abil, mis eristavad kaht tüüpi protsesse sotsiaalse info töötlemisel. Esimene tüüp (Tüüp 1) viitab teadvustamata, kiirele, efektiivsele, automaatsele ja võrdlemisi jäigale infotöötlusele ajus. Teine tüüp (Tüüp 2) viitab aeglasele, paindlikule, enam pingutust nõudvale ja osaliselt teadvustatud infotöötlusele. Artiklis väidetakse, et interaktsionism keskendub peamiselt sotsiaalse tunnetuse neile aspektidele, mille aluseks on esimest tüüpi infotöötlus, samal ajal kui vaimulugemise teooriad (teooria-teooria ja simulatsiooniteooria) keskenduvad pigem neile aspektidele, mis on seotud teist tüüpi infotöötlusega. Kuna need kaks teoreetilist lähenemist keskenduvad sotsiaalset tunnetust võimaldavatele eri tüüpi kognitiivsetele protsessidele, tuleks neid käsitada mitte vastastikku välistavate, vaid teineteist täiendavatena. Hiljutiste duaalsete protsesside neuroteaduslike uuringute analüüsi alusel esitatakse hüpotees, et inimestevahelise interaktsiooni korral aktiveeruvad mõlemat tüüpi infotöötlusprotsessid, mis teineteist mõjutavad. Artiklis esitatud hüpoteesi põhjal töötame Volkswagen Stiftungi poolt *European Platform for European Platform for Life Sciences, Mind Sciences, and the Humanities* raames rahastatud uurimisrühmaga välja

neuroteaduslikke pilooteksperimente. Uurimisrühma kuuluvad lisaks minule filosoofid Marion Godman (Helsingi ülikool, Cambridge'i ülikool) ja Mog Stapleton (Stuttgarti ülikool), ning neuroteadlased Wouter van den Bos (Max Plancki inimarengu instituut Berliinis), Christoph Teufel (Cambridge'i ülikool) ja Marijn van Wingerden (Düsseldorfi Heinrich Heine ülikool). (Vt http://social-interaction.eu/).

Artikkel 2. Teine artikkel "Theory of mind and the unobservability of other minds" (vt lisa 2) valmis koostöös filosoof Nivedita Gangopadhyayga, kes töötab hetkel Bochumi ülikoolis ja see ilmus ajakirjas *Philosophical Explorations*. Artikkel analüüsib interaktsionistlike kriitikute sageli esitatud etteheidet, mille kohaselt tuginevad vaimulugemise teooriad eeldusele, et teiste olendite vaim on vaadeldamatu. Kuivõrd interaktsionismi järgi on osa vaimuseisundeid vahetult tajutavad, siis toetuvad interaktsonistide väitel vaimulugemise teooriad fundamentaalselt ekslikule eeldusele selle kohta, mida sotsiaalse tunnetuse teooriad üldse peaksid seletama. Artiklis näidatakse, et interaktsionistide kriitika põhineb vaimulugemise teooriate vääritimõistmisel ega pea paika. Artikli esimeses osas esitatakse neli võimalikku tõlgendust vaimuseisundite vaadeldamatuse väitele: 1) metafüüsiline tõlgendus (vaim on entiteet, mida ei ole võimalik vaadelda); 2) fenomenoloogiline tõlgendus (vaimuseisundid ei ole kunagi kogemuslikult antud kellelegi peale nende seisundite omaniku); 3) epistemoloogiline tõlgendus (meil ei saa olla vahetut teadmist teiste inimeste vaimuseisunditest); 4) psühholoogiline tõlgendus (vaimuseisundite omistamise aluseks ei saa olla üksnes tajuprotsessid). Põhjendatakse, miks metafüüsiline, epistemoloogiline ega fenomenoloogiline tõlgendus ei rakendu vaimuseisundite omistamise teooriate kontekstis, seega jääb ainsaks võimalikuks tõlgenduseks psühholoogiline väide, mille kohaselt vaimuseisundite omistamiseks tarvilikud psühholoogilised protsessid hõlmavad enamat kui vaid tajuprotsesse. Artikli teises osas võetakse psühholoogiline väide luubi alla ja eristatakse selle kolme versiooni: tugevat (tajuda on võimalik vaid füüsilisi liigutusi), keskmist (taju abil on võimalik eristada subjekte ja objekte, kuid ei ole võimalik omistada konkreetseid vaimuseisundeid) ja nõrka (teatud tüüpi vaimuseisundite, nt uskumuste, omistamiseks on tarvis enamat kui taju, nt teoreetilist järeldamist või simulatsiooni). Lähemal uurimisel selgub, et interaktsionistid omistavad vaimulugemise teooriatele psühholoogilise väite tugevaimat versiooni, kuid viimased eeldavad psühholoogilist vaadeldamatust kõigest keskmises või nõrgas tähenduses. Kogu analüüsist järeldub, et interaktsionistide kriitika vaimulugemise teooriate väidetava vaadeldamatuse eelduse aadressil ei pea paika. Artikli kolmandas osas tuuakse Husserli tajuteooriast lähtuvalt välja oluline tingimus vaimuseisundite vaadeldavusele: nimelt ei ole vaimuseisundid vaadeldavad *samal viisil* nagu füüsiliste objektide sensoorsed omadused. Kuna vaimulugemise teooriad on eeltoodud tingimusega kooskõlas, siis selgub, et interaktsionismil on vaimulugemise teooriatelt midagi olulist õppida: nimelt seda, et on tarvis seletada vaimuseisundite omistamise aluseks olevaid subpersonaalseid protsesse tingimusel, et vaimuseisundid ei ole vaadeldavad samal viisil nagu füüsilised objektid.

Artikkel 3. Kolmas artikkel kannab pealkirja ”We read minds to shape relationships” (vt lisa 3). Artiklis keskendutakse kahele küsimusele: 1) Mis on vaimuseisundite eksplitsiitse omistamise funktsioon inimese sotsiaalses tunnetuses? 2) Millised motivatsioonilised protsessid on aluseks vaimuseisundite omistamisele? Nimetatud küsimusi ei ole vaimulugemise teooriate raamistikus empiiriliselt uuritud, kuivõrd valdavalt on eeldatud, et vaimulugemine on igasuguse sotsiaalse tunnetuse keskne komponent, mis teenib teiste inimeste käitumise ennustamise ja seletamise funktsiooni. Seevastu interaktsionistid väidavad, et vaimulugemine on sotsiaalse tunnetuse jaoks perifeerse tähtsusega, mis samuti ei ole teadlasi motiveerinud vaimulugemise funktsiooni lähemalt uurima. Samas ei ole kumbki debati osapool pakkunud veenvat seletust sellele, miks inimesed teistele vaimuseisundeid omistavad ega vasta küsimustele, mis on vaimulugemise evolutsiooniline funktsioon, miks see on ”vaimulugejale” isiklikult kasulik, või mis motiveerib inimesi teatud olukordades vaimuseisundeid omistama. Nii vaimulugemise teooriad kui interaktsionism on ignoreerinud sotsiaalse tunnetuse üht olulist mõõdet: inimestevahelisi suhteid. Inimestevahelised suhted on meie igapäevaelu oluline osa, mis mõjutab otseselt sotsiaalse interakteerumise viise ning sotsiaalseid tunnetusprotsesse. Artiklis tutvustatakse inimestevaheliste suhete aluseks olevate kognitiivsete protsesside teooriat: Alan Fiske poolt välja töötatud *suhtemudelite teooriat*. Nimetatud teooriast lähtuvalt esitatakse hüpotees, et vaimulugemine võimaldab *sotsiaalseid suhteid reguleerida*: vaimulugemise peamiseks funktsiooniks ei ole mitte teiste inimeste käitumise ennustamine ja seletamine, justkui oleks tegu teaduslikku seletust nõudvate objektidega, vaid inimestevaheliste suhete vastastikune reguleerimine, kuna vaimuseisundite omistamine mõjutab korraga nii tõlgendaja kui tõlgendatava käitumist. Nimetatud hüpotees kirjutatakse lahti empiiriliselt kontrollitavate väidetena. Nimelt võib hüpoteesist lähtuvalt ennustada, et inimesed omistavad vaimuseisundeid tüüpiliselt siis: 1) kui sotsiaalse suhte mudel on ebaselge, kui sotsiaalne interaktsioon muutub problemaatiliseks või selle tulemus on oluline, kuid ebakindel; 2) kui soovitakse olemasoleva sotsiaalse suhte vormi muuta, või usutakse, et partneril on vastav soov; 3) kui on oluline ette näha, kuidas erinevad inimestevahelised suhted üksteist mõjutavad; 4) kui antakse moraalseid hinnaguid teiste käitmisele (kultuurides, kus moraalsete hinnagute puhul on oluline võtta arvesse teole eelnenud kavatsusi ja muid vaimuseisundeid). Väidete testimiseks pakutakse välja *kogemusproovi meetod* (*experience sampling*). Väide, et vaimulugemisel on inimestevaheliste suhete reguleerimise funktsioon ei tähenda aga, et inimesed oleksid selle funktsioonist teadlikud. Eraldi hüpoteesina väidetakse, et vaimulugemise peamiseks vallandajaks võivad olla sotsiaalsed emotsioonid. Artiklis esitatud arutlusest järeldub, et sotsiaalsete suhete tunnetamine on inimestevahelise läbikäimise jaoks fundamentaalsem kui vaimulugemine – viimane on tõenäoliselt evolutsiooniliselt hilisem nähtus kui esimene.

Väitekiri põhjendab, miks oleks tarvis suunduda hetkel aktuaalse interaktsionismi ja vaimulugemise teooriate vahelise debati juurest terviklikuma ja laiahaardelisema käsitluse poole, mis haaraks endasse mõlemad lähenemised.

See võimaldaks lisaks sotsiaalse tunnetuse üksikaspektide uurimisele uurida ka seda, kuidas erinevad sotsiaalse tunnetuse komponendid üheskoos funktsioneerivad. Artiklis 1 esitatud empiiriline hüpotees duaalsete protsesside vastastikmõju kohta on juba aluseks neuroteaduslike eksperimentide väljatöötamisele, ning loodetavasti pälvivad ka artiklis 3 esitatud hüpoteesid vaimulugemise funktsiooni kohta piisavalt huvi empiiriliste teadlaste seas. Tahaks loota, et tulevikus näeme sotsiaalse tunnetuse uurimise olulise osana sotsiaalsete suhete ja nende aluseks olevate tunnetusprotsesside uurimist. Seejuures saavad filosoofid olulisel määral kaasa aidata erinevate teadusharude vaheliste sildade loomisele – iseäranis suur arengupotentsiaal tundub olevat kognitiivpsühholoogia, sotsiaalpsühholoogia ja antropoloogia vahelisel koostööl, mida on seni sotsiaalse tunnetuse uurimisel kahetsusväärselt vähe esinenud.

CURRICULUM VITAE

Name:	Vivian Bohl
Date of Birth:	May 3, 1981
Citizenship:	Estonian
Phone:	+372 737 5314
E-mail:	vivian.bohl@ut.ee, vivian.bohl@gmail.com
Current occupation:	Specialist of the International Curriculum, Institute of Philosophy and Semiotics, University of Tartu

Education

2009–2014	Doctoral studies, University of Tartu
Feb–March 2013	Visiting researcher at University of California, Los Angeles
Jan–June 2012	Visiting researcher at the Centre for Subjectivity Research, Copenhagen
2008	MA in Philosophy, University of Tartu
2005–2006	Visiting student of Philosophy at the University of Konstanz
2004	BA in Philosophy (*cum laude*), University of Tartu
1987–1999	Tartu Mart Reinik Gymnasium

Research Interests:

Philosophy of mind, phenomenology, social cognition, relational models theory

Selection of Publications

Peer-Reviewed Articles Published in International Journals:

Bohl, Vivian. 2014. We read minds to shape relationships. *Philosophical Psychology*, pp. 1–21. Published online: 12 March 2014.

Bohl, Vivian; Gangopadhyay, Nivedita. 2013. Theory of mind and the unobservability of other minds. *Philosophical Explorations*, pp. 1–20. Published online: 30 July 2013.

Bohl, Vivian; van den Bos, Wouter. 2012. Toward an integrative account of social cognition: Marrying theory of mind and interactionism to study the interplay of Type 1 and Type 2 processes. *Frontiers in Human Neuroscience*, pp. 1–15. Published online: 11 October, 2012.

Bohl, Vivian. 2011. Milleks on sotsiaalse tunnetuse uurimisvaldkonnas tarvis filosoofiat? (Why does social cognition research need philosophy?). *Studia Philosophica Estonica*, 4.1, pp. 20–51.

Other Peer-Reviewed Articles:

Bohl, Vivian. 2011. Sotsiaalne tunnetus kui hübriidelevant [Social cognition as a hybrid elephant]. In B. Mölder & J. Kangilaski (Eds.), *Filosoofia ja analüüs: Analüütilise filosoofia seminar 20*. Tartu: EÜS Veljesto, pp. 219–258.

Bohl, Vivian. 2009. Maurice Merleau-Ponty. In E. Annus (Ed.), *20. sajandi mõttevoolud*. Tartu University Press, pp. 263–286.

Book Reviews:

Bohl, Vivian; Fiske, Alan. 2014. Review of *In and Out of Each Other's Bodies: Theory of Mind, Evolution, Truth, and the Nature of the Social*. Maurice Bloch. Boulder, CO: Paradigm, 2012. *American Ethnologist Book Reviews*, *41*(1), pp. 214–215.

Bohl, Vivian. 2013. Omistatud vaim. (Review of *Mind Ascribed: An Elaboration and Defence of Interpretivism.* Bruno Mölder. Amsterdam: John Benjamins, 2010.) *Akadeemia*, 2013/11, pp. 2080–2089.

Bohl, Vivian. 2012. Mina ja minatus ida ja lääne filosoofias [Self and no-self in Eastern and Western Philosophy.] (Review of *Self, No Self? Perspectives from Analytical, Phenomenological, and Indian Traditions*. Eds. M. Siderits, E. Thompson, D. Zahavi. Oxford: Oxford University Press, 2011.) *Sirp*, 10 (3384), pp. 4–5.

Bohl, Vivian. 2010. Tagurpidi Merleau-Ponty [Merleau-Ponty backwards.] (Review of *Nähtav ja nähtamatu*. Maurice Merlau-Ponty. Translated by Mirjam Lepikult. Tallinn: Varrak, 2010.) *Sirp*, 43 (3322), lk 4.

Awards and Scholarships:

2013 UT Graduate School of Linguistics, Philosophy and Semiotics Scholarship for research at UCLA 5.02.2013–22.03.2013

2012 *DoRa* Scholarship for research at the Centre for Subjectivity Research, University of Copenhagen 15.01.2012–15.06.2012

Sept 2011 First prize (graduate level) at the student essay contest of the VII Estonian Annual Conference of Philosophy.

Aug 2010 *DoRa* Scholarship for participation at the Copenhagen Summer School in Phenomenology and Philosophy of Mind, Center for Subjectivity Research, University of Copenhagen, August 2010.

April 2010 *Kristjan Jaak* Scholarship for participation at the VIII Annual Conference of the Nordic Society for Phenomenology, Södertörn University College in Stockholm.

2005–2006 *Herbert Quandt Foundation* Scholarship for studying at the University of Konstanz in 2005–2006.

Participation in Research Projects:

2014– *Disagreements: Philosophical Analysis* (2014–2019). Project nr IUT20-5. Principal Investigator: Margit Sutrop

2012–2015 *Mind Without Mental States?* (2012–2015). Project nr ETF9117. Principal investigator: Dr. Bruno Mölder

2012–2014 *Social Interaction: The Interplay of Pre-reflective and Reflective Processes* supported by the Volkswagen Foundation in the framework of *European Platform for Life Sciences, Mind Sciences, and the Humanities*. Members of the research group: Vivian Bohl,

Wouter van den Bos, Marion Godman, Mog Stapleton, Christoph Teufel, Marijn van Wingerden

2010–2013 *Critical Analysis of Relativism and Pluralism Regarding Truth and Knowledge, Norms and Values* (2008–2013). Project nr SF0180110s08. Principal investigator: Prof. Margit Sutrop

Selection of Presentations

Bohl, V. "We read minds to shape relationships." Poster presentation. Conference: *The Future of Social Cognition: Paradigms, Concepts and Experiments*, Ruhr-University Bochum, June 2014

Bohl, V. "Are mental states unobservable according to theories of mindreading?" Invited oral presentation. Workshop: *The Scope and Limits of Direct Perception*. Center for Subjectivity Research, University of Copenhagen

Bohl, V. "Relational models in the second person perspective." Oral presentation. Conference: *The Second-Person Perspective in Science and the Humanities*. Ian Ramsey Center, University of Oxford, July 2013

Bohl, V. "Towards an integrative account of social cognition." Poster presentation. Summer School: *Embodied Inter-subjectivity: the 1st person and the 2nd person perspective*. Aegina, Greece, June 2013

Bohl, V. "Theory of mind and the unobservability of other minds." Poster presentation. Conference: *Social Understanding: Evolution, Culture and Development*. Ruhr-University Bochum, Sept 2012

Bohl, V. "Are other minds unobservable according to theory of mind?" Oral presentation. Center for Subjectivity Research, University of Copenhagen, June 2012

Bohl, V. "An integrative account of social cognition." Oral presentation. Conference: *Pre-Reflective and Reflective Processes in Social Interaction*. University of Cambridge, March 2012

Bohl, V. "Kas teadus vajab filosoofiat?" ["Does Science Need Philosophy?"] Oral presentation. VII Estonian Annual Conference of Philosophy, University of Tartu, Sept 2011

Bohl, V. "Theory of mind and direct perception of others." Poster presentation. Conference *Embodiment, Intersubjectivity and Psychopathology*, University of Heidelberg, Sept–Oct 2010

Bohl, V. "Theory of mind and direct perception of others." Oral presentation. Summer School: *Phenomenology and Philosophy of Mind,* Center for Subjectivity Research, University of Copenhagen, Aug 2010

Bohl, V. "Kuidas me teisi mõistame?" ["How Do We Understand Others?"] Oral presentation. VII Estonian Annual Conference of Philosophy, Tallinn Technical University, May 2010.

Bohl, V.; Mölder, B. "The directness of the other." Oral presentation. *VIII Annual Conference of the Nordic Society for Phenomenology*, Södertörn University, April 2010

Organisation of Events:

April 2013	*Methods in Studying Social Cognition*, Heinrich-Heine University of Düsseldorf (with W. van den Bos, M. Godman, M. Stapleton, C. Teufel and M. van Wingerden)
March 2012	*Pre-reflective and Reflective Processing in Social Interaction*, Clare College, University of Cambridge (with W. van den Bos, M. Godman, M. Stapleton, C. Teufel and M. van Wingerden)
Aug–Sept 2011	VII Estonian Annual Conference of Philosophy, University of Tartu (main organizer)

Courses Taught:

"Phenomenology and Philosophy of Mind", with Bruno Mölder, University of Tartu, 2009/10 autumn semester.

ELULOOKIRJELDUS

Nimi:	Vivian Bohl
Sünniaeg:	3. mai 1981
Kodakondsus:	Eesti
Telefon:	+372 737 5314
E-post:	vivian.bohl@ut.ee, vivian.bohl@gmail.com
Praegune töökoht:	Rahvusvahelise õppekava spetsialist, Filosoofia ja semiootika instituut, Tartu Ülikool

Haridustee:

2009–2014	doktoriõpingud filosoofias, Tartu Ülikool
2013 veebr–märts	külalisuurija California ülikoolis Los Angeleses
2012 jaan–juuni	külalisuurija Subjektiivsuse uurimiskeskuses Kopenhaagenis
2008	teadusmagistrikraad filosoofias, Tartu Ülikool
2005–2006	külalistudeng Konstanzi ülikoolis
2004	bakalaureusekraad filosoofias (*cum laude*), Tartu Ülikool
1987–1999	Tartu Mart Reiniku Gümnaasium

Peamised uurimisvaldkonnad:
Vaimufilosoofia, fenomenoloogia, sotsiaalne tunnetus, suhtemudelite teooria

Valik publikatsioone:

Teaduslikud artiklid rahvusvahelise levikuga väljaannetes:

Bohl, Vivian. 2014. We read minds to shape relationships. *Philosophical Psychology*, lk 1–21. Avaldatud veebis: 12 märts 2014.

Bohl, Vivian; Gangopadhyay, Nivedita. 2013. Theory of mind and the unobservability of other minds. *Philosophical Explorations*, lk 1–20. Avaldatud veebis: 30 juuni 2013.

Bohl, Vivian; van den Bos, Wouter. 2012. Toward an integrative account of social cognition: Marrying theory of mind and interactionism to study the interplay of Type 1 and Type 2 processes. *Frontiers in Human Neuroscience*, lk 1–15. Avaldatud veebis: 11 oktoober 2012.

Bohl, Vivian. 2011. Milleks on sotsiaalse tunnetuse uurimisvaldkonnas tarvis filosoofiat? *Studia Philosophica Estonica*, 4.1, lk 20–51.

Muud teaduslikud artiklid:

Bohl, Vivian. 2011. Sotsiaalne tunnetus kui hübriidelevant. B. Mölder ja J. Kangilaski (Toim.), *Filosoofia ja analüüs: Analüütilise filosoofia seminar 20*. Tartu: EÜS Veljesto, lk. 219–258.

Bohl, Vivian. 2009. Maurice Merleau-Ponty. E. Annus (Toim.), *20. sajandi mõttevoolud*. Tartu Ülikooli Kirjastus, lk 263–286.

Raamatuarvustused:

Bohl, Vivian; Fiske, Alan. 2014. Arvustus raamatule *In and Out of Each Other's Bodies: Theory of Mind, Evolution, Truth, and the Nature of the Social*.

Maurice Bloch. Boulder, CO: Paradigm, 2012. *American Ethnologist Book Reviews*, *41*(1), lk 214–215.

Bohl, Vivian. 2013. Omistatud vaim. (Arvustus raamatule *Mind Ascribed: An Elaboration and Defence of Interpretivism*. Bruno Mölder. Amsterdam/ Philadelphia: John Benjamins Publishing Company, 2010). *Akadeemia*, 2013/11, lk 2080–2089.

Bohl, Vivian. 2012. Mina ja minatus ida ja lääne filosoofias. (Arvustus kogumikule *Self, No Self? Perspectives from Analytical, Phenomenological, and Indian Traditions*. Toimetanud M. Siderits, E. Thompson, D. Zahavi. Oxford University Press, 2011). *Sirp*, 10 (3384), lk 4–5.

Bohl, Vivian. 2010. Tagurpidi Merleau-Ponty. (Arvustus raamatule *Nähtav ja nähtamatu*. Maurice Merleau-Ponty. Prantsuse keelest tõlkinud Mirjam Lepikult. Tallinn: Varrak, 2010). *Sirp*, 43 (3322), lk 4.

Saadud uurimistoetused ja stipendiumid:

2013 TÜ Keelteaduse, filosoofia ja semiootika doktorikooli stipendium uurimistööks California ülikoolis, Los Angeleses 5.02–22.03.2013

2012 DoRa stipendium uurimistööks Subjektiivsuse uurimiskeskuses, Kopenhaageni ülikoolis 15.02–15.06.2012

2011 sept Eesti filosoofia aastakonverentsi raames korraldatud tudengite esseevõistluse 1. preemia (doktoriõppe tasemel)

2010 aug DoRa stipendium Kopenhaageni Fenomenoloogia ja vaimufilosoofia suvekoolis osalemiseks

2010 apr Kristjan Jaagu stipendium osalemiseks Põhjamaade Fenomenoloogiaühingu VIII aastakonverentsil Södertörni ülikoolis

2005–2006 Herbert Quandt'i fondi stipendium Konstanzi ülikoolis õppimiseks 2005–2006 õppeaastal

Osalemine uurimisprojektides:

2014– *Lahkarvamuste filosoofiline analüüs* (2014–2019). Projekti number: IUT20-5. Vastutav täitja: Margit Sutrop

2012–2015 *Vaim ilma vaimuseisunditeta?* (2012–2015). Projekti number: ETF9117. Vastutav täitja: Bruno Mölder

2012–2014 *Social Interaction: The Interplay of Pre-reflective and Reflective Processes.* Projekti rahastabVolkswagen Stiftung *European Platform for Life Sciences, Mind Sciences, and the Humanities* raames. Uurimisrühma liikmed: Vivian Bohl, Wouter van den Bos, Marion Godman, Mog Stapleton, Christoph Teufel, Marijn van Wingerden.

2010–2013 *Relativismi ja pluralismi kriitiline käsitlus tõe ja teadmise, normide ja vääruste suhtes* (2008–2013). Projekti number: SF0180110s08. Vastutav täitja: Margit Sutrop

Valik ettekandeid:

Bohl, V. "We read minds to shape relationships." Stendiettekanne. Konverents *The Future of Social Cognition: Paradigms, Concepts and Experiments*, Bochumi ülikool, juuni 2014

Bohl, V. “Are mental states unobservable according to theories of mindreading?” Kutsutud suuline ettekanne. Töötuba: *The Scope and Limits of Direct Perception*. Subjektiivsuse uurimiskeskus, Kopenhaageni ülikool, dets 2013

Bohl, V. “Relational models in the second person perspective.” Suuline ettekanne. Konverents *The Second-Person Perspective in Science and the Humanities*. Ian Ramsey keskus, Oxfordi ülikool, juuli 2013

Bohl, V. “Towards an integrative account of social cognition.” Stendiettekanne suvekoolis *Embodied Inter-subjectivity: the 1st person and the 2nd person perspective*. Aegina, Kreeka, juuni 2013

Bohl, V. “Theory of mind and the unobservability of other minds.” Stendiettekanne konverentsil *Social Understanding: Evolution, Culture and Development*. Bochumi ülikool, sept 2012

Bohl, V. “Are other minds unobservable according to theory of mind?” Suuline ettekanne. Subjektiivsuse Uurimiskeskus, Kopenhaageni ülikool, juuni 2012

Bohl, V. “An integrative account of social cognition.” Suuline ettekanne. Konverents *Pre-Reflective and Reflective Processes in Social Interaction*. Cambridge’i ülikool, märts 2012

Bohl, V.” Kas teadus vajab filosoofiat?” Suuline ettekanne VII Eesti filosoofia aastakonverentsil, Tartu Ülikool, sept 2011

Bohl, V. “Theory of mind and direct perception of others.“ Stendiettekanne. Konverents *Embodiment, Intersubjectivity and Psychopathology*, Heidelbergi ülikool, sept–okt 2010.

Bohl, V. “Theory of mind and direct perception of others.“ Suuline ettekanne. Suvekool *Phenomenology and Philosophy of Mind*, Subjektiivsuse uurimiskeskus, Kopenhaageni ülikool, aug 2010

Bohl, V. “Kuidas me teisi mõistame?” Suuline ettekanne Eesti filosoofia VI aastakonverentsil, Tallinna Tehnikaülikool, mai 2010

Bohl, V., Mölder, B. “The directness of the other.“ Suuline ettekanne Põhjamaade Fenomenoloogiaühingu VIII aastakonverentsil, Södertörni kõrgkool, aprill 2010

Konverentside korraldamine:

2013 aprill	*Methods in Studying Social Cognition*, Düsseldorfi Heinrich–Heine ülikool (koos W. van den Bosi, M. Godmani, M. Stapletoni, C. Teufeli ja M. van Wingerdeniga)
2012 märts	*Pre-reflective and Reflective Processing in Social Interaction*, Clare College, Cambridge’i ülikool Suurbritannias (koos W. van den Bosi, M. Godmani, M. Stapletoni, C. Teufeli ja M. van Wingerdeniga)
2011 aug–sept	Eesti filosoofia VII aastakonverentsi peakorraldaja, Tartu Ülikool

Õpetatud kursusi:

“Fenomenoloogia ja vaimufilosoofia” (koos Bruno Mölderiga), Tartu Ülikool, 2009/2010 sügissemestril.

DISSERTATIONES PHILOSOPHICAE UNIVERSITATIS TARTUENSIS

1. **Jüri Eintalu.** The problem of induction: the presuppositions revisited. Tartu, 2001.
2. **Roomet Jakapi.** Berkeley, mysteries, and meaning: a critique of the non-cognitivist interpretation. Tartu, 2002.
3. **Endla Lõhkivi.** The sociology of scientific knowledge: a philosophical perspective. Tartu, 2002.
4. **Kadri Simm.** Benefit-sharing: an inquiry into justification. Tartu, 2005.
5. **Marek Volt.** The epistemic and logical role of definition in the evaluation of art. Tartu, 2007.
6. **Aive Pevkur.** Professional ethics: philosophy and practice. Tartu, 2011.
7. **Toomas Lott.** Plato on Belief (*doxa*) *Theaetetus* 184B–187A. Tartu, 2012, 208 p.
8. **Jaanus Sooväli.** Decision as Heresy. Tartu, 2013, 153 p.
9. **Ave Mets.** Normativity of scientific laws. Tartu, 2013, 217 p.