Levels of Social Embodiment – Towards a Unifying Perspective on Social Cognition

Inauguraldissertation
zur Erlangung des Akademischen Grades
eines Dr. phil.,
vorgelegt dem Fachbereich 05 – Philosophie und Philologie,
der Johannes Gutenberg-Universität
Mainz
von
Lisa Anna Quadt
aus Koblenz
Mainz
2016

Referent:
Ko-Referent:
Tag des Prüfungskolloquiums: 05.10.2016

Contents

List of Figures vi

Acknowledgments vii

Introduction 1

Goals and Structure 3

Summary of Chapters 4

1. Theory-Theory and Simulation-Theory: The Mindreading Debate 4
2. Phenomenology and Enactivism: The ‘Phenactive’ Approach to Being Social 7
3. Social Cognitive Neuroscience: Empirical Findings and Conceptual Challenges . 7
4. Interlude: The State of the Debate on Social Cognition 8
5. Building Block I: First-, Second-, and Third-Order Embodiment 10
6. Building Block II: Predictive Processing 11
7. Towards a Unifying View on Social Cognition: Levels of Social Embodiment .. 12

Part I

1. Theory-Theory and Simulation-Theory: The Mindreading Debate 15

1.1. The Historical Growth of Theory-Theory 16

1.1.1. The Problem of Other Minds 16

1.1.2. The Argument from Analogy 19

1.2. Social Understanding as Theory Construction – The Case of Theory-Theory..... 25

1.2.1. The Genius Jones Invents Theory-Theory 27

1.2.2. The Status of Folk Psychology 31

1.2.3. Non-Human Theory of Mind 34

1.2.4. Empiricist Theory-Theory 35

1.2.5. Nativist Theory-Theory 37

1.2.6. Developmental Psychology and Theory of Mind 39

1.3. The Historical Growth of Simulation-Theory 44

1.3.1. Lipps’ Conception of Empathy 45

1.3.2. Simulation-Theory vs. Theory-Theory 48

1.3.3. Putting Oneself in the Other's Shoes ... 49
1.3.4. Constructing the Other – Simulation as Interpretation ... 52
1.3.4. Mirror Neurons and Low-level Simulation ... 55
1.3.5. Embodied Simulation ... 58
1.4. The Mindreading Debate ... 62
1.4.1. The Common Background of ST and TT ... 63
1.4.2. Going Sub-Personal ... 65
1.5. Evaluating the Debate ... 67
1.5.1. The Terminological Landscape ... 68
1.5.2. A Debate Leading Nowhere?! ... 72
1.5.3. Mindreading – The Central Explanandum for Social Cognition Research? . 73
2. Phenomenology and Enactivism: The 'Phenactivist' Approach to Being Social ... 76
2.1. Traditional Phenomenology ... 77
2.1.1. Direct Perception and Criticism of Previous Views ... 78
2.1.2. The Phenomenological Conception of Empathy ... 80
2.1.3. Embodiment and Intersubjectivity ... 83
2.2. The Modern Phenomenological Perspective ... 86
2.2.1. Intersubjectivity in the Modern Phenomenological Perspective ... 87
2.2.2. Perception: Direct and Smart ... 90
2.2.3. Evaluating the Phenomenological Perspective ... 92
2.3. Making Sense of One Another – Enactivism ... 96
2.3.1. The Embodied Mind ... 97
2.3.2. Participatory Sense-Making ... 103
2.3.3. The Interactive Brain Hypothesis ... 106
2.3.4. The Perceptual Crossing Paradigm ... 107
2.3.5. Evaluation ... 112
3. Social Cognitive Neuroscience: Empirical Findings and Conceptual Challenges ... 120
3.1. The Historical Growth of a Genuine Research Field ... 120
3.1.1. The Rise of Social Neuroscience ... 121
3.1.2. The Emergence of Social Cognitive Neuroscience ... 124
3.2. The 'Social Brain' ... 127
3.2.1. The Social Brain Hypothesis (SBH) ... 128

3.2.2. Networks for Social Cognition ... 129
3.2.3. The Social Perception Network ... 130
3.2.4. The Mentalizing Network ... 134
3.2.5. The Mirror Neuron Network ... 136
3.2.6. The Empathy Network ... 141
3.3. Conceptual Challenges ... 143
3.3.1. How Special is Social Cognition? ... 143
3.3.2. Terminology ... 149

Interlude

4. Interlude: The State of the Debate on Social Cognition ... 155
4.1. A Shift in the (Social) Cognitive Sciences? ... 156
4.1.1. The Pragmatic Turn ... 157
4.1.2. The Interactive Turn ... 158
4.1.3. A Reluctant Turn ... 161
4.2. Pluralistic Accounts of Social Cognition ... 168
4.2.1. The Multiplicity View ... 168
4.2.2. Multiplicity Needs Consistency ... 170

Part II

5. Building Block I: First-, Second-, and Third-order Embodiment (1-3E) ... 176
5.1. Theoretical Basics of 1-3E ... 176
5.1.1. The Theoretical Scope ... 177
5.1.2. First-, Second-, and Third-Order Embodiment ... 179
5.1.3. Transparency and Opacity ... 181
5.2. Discussion ... 182
5.2.1. The Notion of Embodiment ... 183
5.2.2. Combining Non-Representational and Representational Views ... 185
5.2.3. Transparency and Opacity Revisited ... 186

6. Building Block II: Predictive Processing (PP) ... 190
6.1. The Basics ... 191
6.1.1. The Brain’s Task ... 192

6.1.2. Hierarchical Predictive Processing ... 193
6.2. Predictive Processing and the Free-Energy Principle ... 195
6.2.1. Keeping Oneself Alive and Well – The Free-Energy Principle ... 196
6.2.2. Predictive Processing as an Instance of the Free-Energy Principle ... 197
6.3. Active Inference ... 198
6.3.1. Minimizing Prediction Error by Acting ... 199
6.3.2. Estimating Uncertainty – The Role of Precision Weighting ... 200
6.3.3. A New View on Motor Control ... 203
6.4. PP and Interoception ... 206
6.4.1. The Cybernetic Brain ... 206
6.4.2. Interoceptive Predictive Processing ... 207
6.5. A New View of the Mind ... 211
6.5.1. Counterfactually-Equipped Representations ... 212
6.5.2. The War on Representations ... 213
6.5.3. Phenactive Predictive Processing? ... 215
6.5.4. Merging 1-3E and PP ... 217
7. Towards a Unifying View on Social Cognition: Levels of Social Embodiment ... 221
7.1. The Basics of 1-3sE ... 221
7.1.1. Purpose and Scope ... 222
7.1.2. A Brief Description of Levels of Social Embodiment ... 224
7.2. First-order Social Embodiment (1sE) ... 226
7.2.1. Embodied Social Inference (EmSI) ... 227
7.2.2. The 'Social Brain' ... 234
7.2.3. Interactive Inference (InI) ... 240
7.3. Second-Order Social Embodiment (2sE) ... 247
7.3.1. The Shared Body Model ... 248
7.3.2. Sharing Representations – The Case of Mirror Neurons ... 250
7.3.3. Problems with Shared Representations ... 254
7.3.4. Contextual Clothing ... 256
7.3.5. Are Others a Mere Version of Ourselves? ... 261
7.3.6. Interactive Inference 2.0 ... 265
7.4. Third-Order Social Embodiment ... 267

7.4.1. The Phenomenology of Social Encounters 268
7.4.2. PPSMC, Transparency and Opacity 271
7.4.3. The Spectrum of Mental Presence 275
7.4.4. Taking Phenomenal Experience to Another Level – The Case of 3sE+ 278

Conclusion 282

Results 282

Future Research 286

Bibliography 289

Deutsche Zusammenfassung 311

Erklärungen gemäß § 6 der Promotionsordnung 313

List of Figures

Fig. 1 The hierarchical reading ... 178

Fig. 2 The systems reading ... 179

Fig. 3 Revised hierarchical reading ... 188

Fig. 4 Active inference ... 205

Fig. 5 Interoceptive predictive processing ... 210

Fig. 6 Action observation ... 252

Fig. 7 The framework of 1-3sE ... 280

Introduction

JULIA: But you know me, Danton!
DANTON: I know your dark eyes, your curly hair, your fine skin and that you call me 'darling Georges'. But! (He points to his forehead and eyes.) Here, here, what lies behind here? Our senses are crude. We'd have to crack open our skulls to know each other, tear out each other's thoughts from the fiber of the brain.
(Danton's Death, Georg Büchner)

How do human individuals understand each other? Do we really have to crack open the skull of another person to truly know her and her intentions? Do we not have the very different experience that without much effort, we feel the sorrow and share the joy of our fellow humans? These are only some of the questions that the research field of social cognition and this work is concerned with. Among those that are looking for answers are philosophers, psychologists, cognitive scientists, neuroscientists, and phenomenologists, forming an ever-growing, diverse field of interest. They are concerned with the cognitive and behavioral processes that make us humans social, that enable us to understand each other, take each other's perspective, jointly achieve a goal, or have empathy for one another. This thesis aims at contributing a philosophical perspective on central questions of the field.[1]

It seems that there are two competing intuitions about how social understanding is brought forth. On the one hand, it appears so simple and easy to grasp the feelings and intentions of another person. When I enter the room and see my friend weeping, I immediately 'just know' that she is sad. It is obvious from her behavior, no skulls cracked open. On the other hand, however, her behavior is really all I have. And behavior, as we all know, can be deceitful and inaccurate. What about her underlying intentions and motivations? How can we access these hidden states of the other person?

The research field on social cognition now appears divided by these intuitions. While some argue that we can effortlessly and directly perceive the contents of another person's mind, others are convinced that mental states need to be inferred, since they cannot be grasped in an immediate manner. Behind these intuitions stand two theoretical camps that are not only divided by their claims about social cognition, but the mind in general. These are enactivist

[1] In this work I will focus on human social cognition.

and phenomenological (henceforth, *phenactivist*), and cognitivist theories that hold crucially distinct assumptions about the metaphysical nature of cognition, and social cognition specifically. Cognitivist theories of social cognition basically assume that our so-called mindreading skills exhaust our social cognitive processes, thus focusing particularly on the brain of individuals. Mindreading has been described as the ability to infer the mental states of others by either forming a theory about them or by simulating their current states. Phenactive views, on the other hand, assume that only in the embodied interaction of two or more agents does social cognition emerge, thereby denying the need for inferential processes.

Given this theoretical diversity and competition, this thesis is mainly motivated by the following question:

> What kind of philosophical theory is needed in order to capture the diverse nature of social cognition in a unified and comprehensive manner?

Working towards an answer will involve examining already existing theories and evaluating their pros and cons. On this basis, we can then ask what is additionally needed to theoretically capture the manifold phenomenon of social cognition.

I am also convinced that this matter needs to be approached in an interdisciplinary manner, taking into account philosophical, phenomenological, psychological and neuroscientific research. For I think that philosophy – in order to yield a comprehensive theoretical framework of the phenomenon – has to be empirically informed. There have been a number of cases in the history of researching social cognition in which new empirical findings have influenced theoretical advances. The most compelling example of such a case is the discovery of so-called mirror neurons, which fire both when an agent executes and merely observes an action (Rizzolatti & Craighero, 2004). After researchers found out about the properties of this specific group of neurons, both philosophers and neuroscientists changed their perspectives on how social cognition is to be framed theoretically. The hitherto prevalent scheme – theory-theory – seemed disproven, while simulation-theory appeared to be empirically evident. On the other hand, there is a lot that philosophy can do for the empirical side of the research field. Not only does social cognitive neuroscience come with very specific conceptual challenges, as I will detail in this thesis. Philosophy can also contribute in giving a meta-perspective on both empirical and theoretical investigations in that it integrates views from several directions. This is what I aim to do in this work –

approach social cognition from a meta-perspective and examine the diverse and manifold findings from several disciplines. Only when we consider both the theoretical and empirical side of a phenomenon, and only when science and philosophy work in tandem can be begin to build an encompassing theoretical framework.

Goals and Structure

The main epistemic goal of this pursuit is twofold. My first aim is to describe traditional and modern approaches to social cognition in more detail and work out both their shortcomings and advantages. I will argue that there is a deep division between different accounts that has theoretical and practical consequences, leaving the current state of debate unable to provide a comprehensive theory of social cognition. Secondly, I suggest ways in which we can start to overcome this divide and begin to move towards a unified account of social cognition. It will be claimed that two theories are especially well-equipped to provide building blocks for such a new account. These two accounts are then merged and applied to social cognition. The structure of this work mirrors these goals and is partitioned in two parts respectively.

Throughout Part I (chapters 1-3) of this thesis, I will distill functional components of social cognition and desiderata for a theory of the phenomenon that shall be considered in Part II. In doing so, three main areas of the research field are scrutinized, namely cognitivism, phenactivism, and social cognitive neuroscience. I thereby gather insights from a range of perspectives that shall provide a critical set of the most relevant components of our target phenomenon that need to be taken into account, and requirements which must be met by a theory that strives to do so. Let me briefly elaborate on the difference between components and desiderata. The latter refers to requirements for the theoretical framework itself. For example, a theory of social cognition is entitled to operate on a consistent set of background assumptions. The desideratum, in this case, would thus be *consistency*. Components, on the other hand, can be seen as elements of social cognition (e.g., interaction is seen as one aspect of social cognition and should thus be taken into account theoretically. These elements can be necessary for social cognition to occur, they can be a constitutive part of the phenomenon, or elements that enable social cognition. There is, of course, a strong relation between components and desiderata, in the sense that each component should be considered and taken into account by the theory.

The Interlude (chapter 4) provides an overview of the state of the debate in the research field on social cognition. Taking together the findings of Part I and current trends in the field, I claim that there are two main problems which hinder a comprehensive view on our target phenomenon. Firstly, all of the theoretical strands I discuss and present have fundamental flaws which leave them unfit for yielding a basis for a coherent perspective on social cognition. This has led some researchers and philosophers to try and combine theoretical elements from the accounts at hand, trying to widen the scope. I claim, however, that such a combination does not come easy, since the elements that are put together are based on contradictory background assumptions. The goal thus must be, or so I propose, to find a theoretical framework that is based on consistent assumptions, but is still open for integrating diverse elements from all theoretical perspectives.

In Part II (chapters 5-7), I set out to find such a framework. I claim that the theory of first-, second-, and third-order embodiment (1-3E; Metzinger, 2014a) and the scheme of predictive processing (PP; Clark, 2016; Hohwy, 2013) provide a set of conceptual tools that is fit to serve this goal. Putting these two building blocks together, I aim to start building a theoretical framework of social cognition that includes the components and fulfills the desiderata that have been worked out in Part I. I will call this framework ‘first-, second-, and third-order social embodiment’ (1-3sE).

Although it is claimed that 1-3sE can help to form a comprehensive view on social cognition, it can only provide a starting point to do so. In future work and with growing empirical evidence, both desiderata and components will need refinement. Along with this, 1-3sE will be in need of modification in order to stay up to date with both theoretical and empirical research.

Summary of Chapters

1. Theory-Theory and Simulation-Theory: The Mindreading Debate

The first chapter serves to give an overview of a debate on social cognition that emerged in the second half of the 20th century and still goes on: the mindreading debate. Its contributions to the research field shall be evaluated and I will argue that the debate has not succeeded in giving a coherent terminological and conceptual approach for social cognition.

I will start with a depiction of historical precursors of mindreading approaches and show that the intuitions that led to the formation of the two main theories in the debate – theory-theory

(TT) and simulation-theory (ST) – go back as far as Descartes and Lipps. From these precursors, we can already derive several components of social cognition.

c_1 – inference
The component of inference results from the assumption that mental states of the other person are hidden causes of perceivable behavior and thus need to be inferred.

c_2 – similarity
If social understanding partially draws on the application of self-related processes to another person, a specific degree of similarity between individuals is needed.

c_3 – self-other distinction
Social cognition requires the distinction between self and other so not to confuse self- and other-related processes.

TT was formulated within the more general discussion on the nature of the human mind by Sellars (1956/1997). Since the prevalent opinion on the matter entailed describing the mind as a linguistic device whose main operation is the manipulation of symbols, it was not far-fetched that social cognition also relies on an equally language-based process. Sellars proposed that in order to infer the contents of another person's mind, we form theories on the basis of folk psychological laws. TT was modified many times and now comes in different versions, the most relevant of which I will depict. Interestingly, a still-prevalent empirical paradigm was derived from its assumptions and has since been used to evaluate whether or not an individual possesses a so-called 'theory of mind' (ToM); the false-belief task (Wimmer & Perner, 1983).

The main opponent of TT was introduced in the 1980s, when psychologists and philosophers started to doubt that the basis for understanding each other are folk psychological rules. Simulation-theory counters that instead of relying on these laws, humans use their own experience and apply it to the other person in order to understand her. This process has been called simulation and is still widely debated today. When mirror neurons were discovered, it was thought that the neural basis for simulation was discovered and that TT was defeated. It turned out, however, that philosophers and scientists soon became inclined to thinking that in order to fully describe social cognition, both simulation and theorizing would be needed. This development of forming hybrid theories is claimed to be possible because TT and ST are based upon very similar background assumptions of the human mind. Investigating these assumptions and the more specific claims of the theories, the following components of social cognition are derived:

c_4 – construction
Construction refers to the ability to consciously construct theories about the mental states of other people.

c_5 – experiential quality
Experiential quality refers to the assumption that social encounters come with different phenomenal signatures that serve a specific function.

c_6 – replication
Replication describes social cognitive processes that involve reproducing the state of the other person in order to gather information about her.

c_7 – embodiment
Embodiment refers to the assumption that the body plays a crucial role for most social cognitive processes.

After having distilled the fruitful contributions of TT and ST, I critically assess the mindreading debate. During this assessment, these desiderata for a philosophical framework of social cognition are gathered:

d_1 – specificity
Specificity postulates the need for the assumption that there are properties of social cognition differentiating it from general cognition.

d_2 – interdisciplinarity
Interdisciplinarity demands that a framework on social cognition considers theoretical and empirical findings from several disciplines and enables a dialogue between them.

d_3 – multi-level analysis
Multi-level analysis asks for a description of social cognition at several levels of analysis in order to capture its diverse components.

d_4 – terminological consistency
Terminological consistency postulates the need for expressive terms that are used in an unequivocal manner.

It will be claimed that the mindreading debate has failed to yield a coherent taxonomy of concepts that would be useful for an interdisciplinary dialogue and that its scope is too narrow to capture the diverse nature of social cognition.

2. Phenomenology and Enactivism: The 'Phenactive' Approach to Being Social

After casting doubts on the mindreading debate, the current main competitor – the phenactive approach to social cognition – is scrutinized and evaluated. It will be argued that there are conceptual and empirical shortcomings that leave this account unfit for yielding a comprehensive perspective on the phenomenon.

Both phenomenological and enactive accounts are described in more detail in order to evaluate their contributions for a theory of social cognition. For phenomenology, it is questioned whether the theory is plausible from an empirical perspective, which is why this desideratum is introduced:

d_5 – empirical plausibility
Empirical plausibility demands that the empirical predictions a theory makes are consistent with actual empirical findings.

Enactivism faces a similar criticism in that it is claimed to be questionable whether the empirical results that are thought to back up the theory are unambiguous. It will be argued that the few studies which are supposed to show that interaction patterns indeed constitute social cognition can also be framed within a non-phenactive picture. Additionally, the terminological landscape of phenactivism appears rather incoherent. On these grounds, phenactivism is also rejected as providing the basis for a comprehensive and coherent account on social cognition. However, the following component of social cognition that has often been neglected by mindreading approaches shall be added to our list:

c_8 – interaction
Interaction is seen as an enabling and contextual component of social cognitive processing.

3. Social Cognitive Neuroscience: Empirical Findings and Conceptual Challenges

In this chapter, the research field of social cognitive neuroscience (SCN) is depicted as a genuinely interdisciplinary one, having grown out of social psychology, cognitive science, and neuroscience. After having described the development of the field, I will sketch the main empirical findings of the investigations of the so-called 'social brain'. The social brain is thought to encompass the regions that underlie social cognition, and will be divided in the following networks:

- social perception network
- mentalizing network
- mirror neuron network
- empathy network

SCN not only comes with a set of theoretical assumptions, but also faces conceptual challenges that will be detailed in the chapter. In focus here is the question of whether or not social cognition is different from general cognition, and if so in which way. Another big challenge for the interdisciplinary research field is to find a terminology that enables a dialogue between the respective disciplines and forms a common ground for conversation. I will address both of these problems and make suggestions on how a philosophical perspective can alleviate them.

4. Interlude: The State of the Debate on Social Cognition

This chapter is thought to put together the findings of the previous sections and thereby provide an overview of the current state of debate in the research field on social cognition. The question I will pursue is whether there is any account available that is able to include all components of social cognition and fulfill the desiderata that have been introduced so far. It will be argued that this is not the case and a consistent philosophical theory on social cognition is still missing.

The influence of phenactive theories has recently caused an interesting movement – the so-called interactive turn (Overgaard & Michael, 2013) – which finds its counterpart in general cognitive science (the pragmatic turn; Engel et al., 2013). The interactive turn describes a supposed paradigm shift in the research field, away from cognitivist and towards phenactive theories. This shift not only has theoretical, but also empirical consequences, in that it dictates to integrate more 'ecologically valid' paradigms and include interactive, embodied, and also emotional processes.

There are three claims that I wish to make concerning the interactive turn. Firstly, I take it that this movement holds exciting possibilities for theoretical and empirical research, since it widens the scope and emphasizes hitherto neglected components of social cognition. This is why a last element of the phenomenon is added:

c_9 – emotions
Social cognitive processes are often influenced by and influence emotional processes.

Secondly, it is questioned whether there is an actual full turn towards phenactivism. Much rather, or so I argue, there is a tendency in the field to pay more attention to phenactive theories and their focus on embodied interaction that does not come without reluctance. This reluctance is asserted to stem from the unwillingness to accept the radical assumptions of phenactivism and the practical and theoretical consequences they bring with them. Consequent upon both the advantages of the interactive turn and the reluctance to fully turn towards phenactivism is my third claim, namely that we need a theoretical framework that integrates insights and components from both cognitivist and phenactivist accounts.

Such an integrative theory, however, does not come without challenges. There have indeed been attempts, especially in philosophy, to combine phenactive and cognitive elements into pluralistic views on social cognition. After presenting one specific attempt – the multiplicity view (Newen, 2015) – I specify the worry that a simple combination of both theories results in an incoherent view on social cognition. For cognitivism and phenactivism come with contradictory background assumptions, it is not possible to put them together without paying attention to the metaphysical demands of each theory. This is why the following desideratum is introduced:

d_6 – consistency
Consistency demands that a theory is built upon non-contradictory background assumptions, but rests on coherent grounds.

The chapter is concluded by stressing that a theory of social cognition needs to encompass a range of components and integrate quite diverse perspectives. To capture this, two more desiderata are added to the list:

d_7 – comprehension
Comprehension asks for encompassing the relevant components of social cognition and comprising elements from several accounts.

d_8 – integration
Integration postulates that a philosophical theory of social cognition must be able to include elements from a variety of interdisciplinary accounts on the phenomenon.

The task is now to start finding a theoretical framework that is able to take on these challenges.

5. Building Block I: First-, Second-, and Third-Order Embodiment

The theory of first-, second-, and third-order embodiment (1-3E; Metzinger, 2014) is the first building block for the framework I am going to propose. 1-3E has been introduced as a theory that serves to describe how phenomenal properties are computationally and physically grounded. The different levels of embodiment are used as levels of description at which a phenomenon can be analyzed. In this sense, the first level of embodiment (1E) can be seen as the physical grounding of a phenomenon, and the second one (2E) as describing the relevant computations. The third level of embodiment (3E) then depicts the phenomenal signature of the target phenomenon. In its original formulation, 1-3E addresses the problem of how the experience of being a self – phenomenal selfhood – is brought forth and aims to describe its computational and physical counterparts.

It is assumed that the three levels are related and build upon each other. Every conscious organism that possesses 3E thus is also in possession of 1E and 2E, since higher levels of embodiment are grounded in its lower counterparts. This enables a comprehensive analysis of a given phenomenon, taking into account its phenomenal properties, its computational processes, and also its physical grounds.

There are several features of the theory that are of particular importance both for merging it with PP, as well as for an application to social cognition. First, its hierarchical structure allows a fruitful combination with predictive processing, as will be described in chapter 6. Further, since it offers a framework for describing a phenomenon at different levels, it is able to integrate a range of components of social cognition. It thus allows to depict social cognition as embodied, interactive, while at the same time focusing on the phenomenal experience that come with social encounters. At the level of 3E, Metzinger distinguishes two properties of phenomenal states: transparency and opacity. While the latter describes states during which individuals are aware of the fact that they are indeed representing something, the former refers to states in which one is fully immersed in the experience. This distinction will be used and extended for describing the phenomenology of social encounters in chapter 6. Additionally, I will point to some shortcomings of 1-3E and show how they can be alleviated.

6. Building Block II: Predictive Processing

Predictive processing is a growingly prominent theory on how action, cognition and perception work. The theory draws a picture of the mind as both embodied and representational, thus being able to integrate many high- and low-level phenomena. It is especially well-equipped to integrate both phenactivist and cognitivist claims, but remains on metaphysically coherent grounds. These features make it a perfect basis for a philosophical theory of social cognition.

The basic assumption of PP is that the brain cannot directly access states of the external world and thus needs to infer the hidden causes of the effects it gets. It does so, or so it is claimed, by predicting its own most likely next state and by then comparing this prediction to the actual incoming sensory signal. On the basis of this comparison, an error signal (prediction error) is generated and used to update the prediction. Prediction error minimization is thought to be the fundamental task of the brain in order to enable a successful navigation of the organism through its environment.

Although this appears *prima facie* as a rather brain-bound cognitivist view on the mind, PP is able to integrate important claims from the phenactive side of the spectrum and depicts a deep connection between perception, action, and cognition. For one, PP is described as an instance of the free-energy principle (FEP; e.g., Friston, 2009) and thereby relates cognition to the biology of organisms. Secondly, prediction error minimization is not only achieved by an internal update of predictive models, but also by engaging in embodied interaction with the environment. This process is called *active inference* and plays a central role in the scheme. Lastly, PP has been related to steering interoceptive processes in order to maintain the homeostasis of the system. *Interoceptive predictive processing* (IPP; Seth, 2013) describes the process of predicting bodily states and minimizing prediction error via the engagement of autonomic reflexes.

In this chapter, I describe PP in more detail and then proceed to discuss how it enables a fresh view on the mind that integrates insights from both cognitivist and phenactivist perspectives. Additionally, it is claimed that PP is not only compatible with 1-3E, but that the merging of these theories will provide the means for a comprehensive framework for social cognition.

7. Towards a Unifying View on Social Cognition: Levels of Social Embodiment

In this last chapter, I introduce the framework of 1-3sE and claim that it can be used as a starting point to integrate the components and fulfill the desiderata that have been worked out in Part I. After describing the scope and purpose of 1-3sE, I depict every level of embodiment in more detail. The basic idea is that at each level of social embodiment, specific aspects can be described from a PP perspective.

The first level (1sE) describes the bodily and neural basis for social cognition. I introduce several conceptual tools that shall serve to depict the phenomenon as embodied and interactive, while at the same time paying attention to the relevant brain mechanisms.

2sE describes the relevant computations and higher-level representational processes that are thought to underlie social understanding. I focus on the notion of 'shared representations', i.e., representations which are used for both self- and other-related processing and discuss several problems that arise when taking a closer look. I claim that with the help of PP, these problems can be alleviated and a redefined notion of shared representations can be formed.

At the third level of social embodiment (3sE), the phenomenal signature of social encounters will be depicted. It will be argued that there is a range of diverse experiences that can be distinguished by locating them on a spectrum of transparent and opaque states. Additionally, the notion of 'counterfactually equipped representations' and the theory of predictive processing of sensorimotor contingencies (PPMSC; Seth, 2014) will be applied to social cognition. With the help of these theoretical tools, a clearer notion of what differentiates transparent form opaque states can be derived. In a last step, I introduce an additional level of social embodiment, 3sE+. This level serves to describe opaque phenomenal states which have the additional signature of representing that one is currently representing something. Since this is not only a very rare, but also a rather specific phenomenal experience, I take it that it deserves a proper level of description. This distinction helps to generate a more nuanced picture on how we experience social situations.

With 1-3sE, I hope to yield a starting point for a philosophical theory of social cognition that is able to capture the diversity of social cognition, while at the same time providing a unifying perspective.

Part 1

1. Theory-Theory and Simulation-Theory: The Mindreading Debate

> *But then if I look out of the window and see men crossing the square, as I just happened to have done, I normally say that I see the men themselves, just as I say I see the wax. Yet do I see any more than hats and coats which could conceal automatons? (Descartes 1641/1996, p. 21)*

How am I supposed to know what you think if I cannot access your mind in the same way that I access mine? How can I be sure that you even have a mind as I do? These were the 'problems of other minds' that arose when Descartes presented his famous methodological doubt and which follow when it is assumed that the mind is only accessible through a privileged first-person perspective. Although a Cartesian notion of the mind was soon discarded, the conviction that mental states are unobservable – if they are not uttered verbally – remained. But then, if I cannot directly see your mental content, how do I come to interpret your behavior?

In this chapter, I aim to present the debate that pivots on these issues. Except for the occasional philosopher addressing the problem of other minds (chapter 1.1.), the core discussion got off around the second half of the 20th century. In focus was the question whether we apply folk psychological laws to other people, or simulate them on the basis of what we know about our own minds and thereby 'read the minds' of others.

The main goal will be to evaluate the so-called mindreading debate, and to argue that it has not succeeded in yielding a useful conceptual framework in which further discussion can be embedded in. In achieving this goal, I will first present the historical growth of the two main theoretical strands which shaped the debate profoundly (chapters 1.2. and 1.3.): theory-theory (TT) and simulation-theory (ST). The idea that we form a theory in order to understand other people was introduced by philosopher Wilfrid Sellars and picked up by philosophers as well as psychologists. The theory was challenged when ST emerged to offer an alternative explanation of how mental states are attributed to other people, namely through a simulation process.

After depicting the development and different versions of ST and TT, I describe the further development of the debate in chapter 1.4., which at its later stages started to shift from either-

or-questions to forming hybrids of simulation and theorizing. I claim that this movement was possible because both theories share fundamental metaphysical background assumptions about how the mind and social cognition works.

In a last step, I evaluate the results of the mindreading debate concerning its terminological and content-related efficiency.

1.1. The Historical Growth of Theory-Theory

The main claim in this section is that some fundamental assertions that make present-day researchers conceive of social cognition as a problematic phenomenon stem from classical discussions in philosophy of mind. In arguing for this statement, I take two steps:

(1) In section 1.1.1, I describe the *problem of other minds,* which results from a Cartesian conception of mind and body. I further distinguish between the *epistemological*, *conceptual*, and *metaphysical problem of other minds* and show how they are related to current discussions of social understanding. In particular, I claim that Descartes' substance dualism led to the description of components of social cognition that are still vividly discussed today.

(2) Chapter 1.1.2. delineates the work of John Stuart Mill as the locus classicus of the argument from analogy. Although the argument has faced profound criticism which leaves it, as I will show, ill-founded, three important components of social understanding can be distilled in discussing the argument from analogy: inference, similarity, and self-other distinction.

1.1.1. The Problem of Other Minds

What is tackled by philosophers caring about the problem of other minds is the strong intuition that fellow human beings have minds just like mine. We are convinced that the people surrounding us act (most of the time) intentionally and rationally and that their (mostly) intelligent behavior depends upon some kind of mental abilities. The assumption that other minds exist is so fundamental that most people – especially non-philosophers – would not even think about it at all. But, as Avramides (2001) points out, if we can doubt that other non-human animals have minds, why not be unsure about whether human animals have them? Therefore, the general problem of other minds is spelled out as "[…] the problem of how to justify the almost universal belief that others have minds very like our own" (Hyslop, 2010). Furthermore, this problem comes in different flavors. The *epistemological problem of other minds* deals with how we come to believe that others have minds and how we gain that knowledge. The *conceptual problem of other minds* asks how we come to a conception of the mind in general and also how we are able to form concepts about other minds. These questions are intertwined, as Avramides (2001, p. 219) points out:

> […] what puts us in a position to so much as raise these questions about the mind of another? In order to raise these epistemological questions I must have a quite general concept of mind – a concept that applies to others as well as myself.

We can further distinguish what I dub the *metaphysical problem of other minds*, which challenges the very basic intuition that other minds exist. Let me now describe these versions of the problem in more detail.

Like many other problems in philosophy of mind that continue to be discussed today, the problem of other minds can be traced back to the work of René Descartes. In his *Meditations* (1641/1986), he describes how his *methodological doubt* forces him to question everything, including not only his own, but also other people's existence. It is famously known that he comes to the conclusion that the only thing he can be sure of is his own thinking, him being someone who is capable of thought. He thus argues for what is now known as *substance dualism*. There must be a *res cogitans*, a non-physical entity also known as the soul, and a *res extensa*, a physical entity, the body. It follows that the existence of anything outside res cogitans, be it my own body, the book in my hands or the person I see walking down the street, can be doubted. Although not explicitly mentioned, the problem of other minds follows from this dualistic view. How can we ever be sure whether other people and their minds exist,[1] if the only thing whose existence is certain is *my own* mind? This is the metaphysical problem of other minds. Neither philosophers in the Cartesian tradition nor those who reject his views have actually acknowledged the problem, often describing it as "the so-called problem of other minds" (Avramides, 2001, pp. 9–11). What is more important and has also gained a lot more attention is what follows from the Cartesian conception of the mind and body. If the only thing I can be sure of and access in a direct

[1] In the sixth meditation, Descartes (1641/1986) tries to avoid a solipsistic view by proving that God exists. Since God is not a deceiver, the external world must exist, rather than being an illusion or dream.

fashion is my own mind, it becomes clear that even if I assume that other minds exist, I can never *access* them (as immediately as my own mind). Thus, despite the (seemingly) avoided solipsism, there is an asymmetry of how to access my own and other minds, prompting the question of how we gain knowledge about other minds. This is the epistemological problem of other minds, as Fulford (2006, p. 741) puts it:

> How can one have knowledge of, or even access in some weaker sense to, other people's minds? If mental states are elements in a private theatre with only one spectator, how can others gain access to them? This is the problem of Other Minds.

What we find here is an important element of social understanding that has ever since made researchers marvel at social abilities. It appears that there is an asymmetry of access between oneself and others. Throughout the course of this thesis, we will keep coming back to this and see that it is not only still pervasive today, but that it should also be seen as a crucial element for social cognition as a mechanism. For now, however, it is important to note is the following.

Although the asymmetry of access assumption *resulted from* Cartesian dualism, it does *not depend on* it. To see this, consider that the basic claim at stake boils down to the fact that there is some kind of input that cannot be given by another person – at least not without fancy future technology. For example, interoceptive input yielding information about my blood-sugar levels is only available for my system and can only be made available by my system. In this somewhat trivial sense, asymmetry of access holds, even from a Non-Cartesian perspective. Put differently, in order to make the claim that there is a difference in how information about self and other is gathered, it is not necessary to assume that the only thing I can be certain of is the existence of my own res cogitans which is a different substance than my res extensa.

Another assumption is implied in this line of thought, emphasized by John Locke and still held by many researchers today, viz. that minds are invisible and unobservable.[2] Locke argues that, since minds are not (at least not *directly*) perceivable, and since we can only have *opinions* about other minds rather than *knowledge,*[3] the knowledge of existence of other

[2] As we will see, phenomenologists deny this assumption vehemently. In their view, which I depict in chapter 2.2., human beings and their minds are fundamentally embodied. They thus claim that we are able to directly perceive mental states in the bodily behavior of others (Scheler, 1912/1973).

[3] For a description of the difference between knowledge and opinion in Locke's work, see Avramides, 2001.

minds remains a probability. By observing other people talk and act, one can *infer* that these people have minds, since the utterance of words is a sign of utterance of ideas (cf. Avramides, 2001, p. 110). At this point, one can find a hint to the idea spelled out by Mill (1872) in greater detail, namely that there is a connection between visible behavior and invisible mental activities that builds the basis for solving the epistemological problem. What Locke and most other philosophers dealing with those problems before the 20th century[4] share is the Cartesian framework they – at least implicitly – ground their work in. It has been Thomas Reid who pointed out that as long as one does not reject Cartesian Skepticism, the problem of other minds can never be resolved. In Reid's view, his opponents take Skepticism too seriously. He argues that this flawed starting point defects the whole system. Although Reid fails to give a positive solution to the problem, he was the first to acknowledge that the issue lies in the Cartesian framework and also the first who explicitly focused on the problem of other minds. It was John Stuart Mill who refused Reid's attempts and claimed that the *argument from analogy* obviates the need to reject the Cartesian framework (Avramides, 2001, p. 212).

1.1.2. The Argument from Analogy

The question that arises from the Cartesian framework is the following: "How do we know whether there is a mind connected with any body that we may encounter?" (Hyslop, 2010). This is what the argument from analogy tries to give an answer to. Famous proponents of this argument are A.J. Ayer, Bertrand Russell, William James and John Stuart Mill. These philosophers found different ways to spell out the argument of analogy, but I shall concentrate on the locus classicus,[5] John Stuart Mill's (1872, pp. 243–244) phrasing of the argument:

> By what evidence do I know, or by what considerations am I led to believe, that there exist other sentient creatures; that the walking and speaking of figures which I see and hear, have sensations and thoughts, or in other words, possess Minds? […] I conclude that other human beings have feelings like me, because, first, they have bodies like me, which I know, in my own case, to be the antecedent condition of feelings; and because, secondly, they exhibit the

[4] E.g., Bishop Berkeley and Nicolas Malebranche. For a review of their work on the problem of other minds, see Avramides, 2001.

[5] There are several reasons why I made this decision. First of all, I will focus on Theodor Lipps' criticism of the argument from analogy as spelled out by Mill in chapter 2.1. Secondly, since the intention of laying out the argument is to describe the basic idea, and not to discuss the different flavors, I deem it sufficient to focus on the classical description of the argument.

> acts and other outward signs, which in my own case I know by experience to be caused by feelings.

Beyond the aspect that follows from the Cartesian framework and that I have depicted in the previous section (asymmetry of access), three premises of the argument are implied in the quotation. The first states that there is a causal link between bodily behavior and the mind. This becomes apparent in Mill's statement that feelings cause "outward signs" and that I know of this causal link from the knowledge I hold about myself and which I gather through experience. That this link is hidden and not perceptually available, but that it is a 'conclusion' that leads to the knowledge of other minds, becomes obvious in the second presupposition: The only thing I can observe of the other is her bodily behavior. The assumption is already implicit in claiming that minds are invisible, but here, the (non-existent) expressiveness of the body is emphasized. The third premise holds that there is a profound psychological similarity between myself and others which is needed in order to make analogy work. To see this, consider the following quotation:

> In general terms, to argue by analogy is to argue on the principle that if a given phenomenon A has been found to be associated with another phenomenon B, then any phenomenon similar to A is very likely to be associated with a phenomenon similar to B. (Borchert, 2006, p. 51)

Thus, analogy only works if humans are indeed psychologically similar and one can apply the knowledge one has about herself to other persons. In sum, Mill argues that only by drawing an analogy from my own case can I ever know that other people have minds like me. Although the questions in current research concentrate on asking how we access the content of other minds rather than asking the classical epistemological question I have just described, the intuition that we do so by inferring mental states from the information we gain by observing the bodily behavior of others, remains. Despite the fact that there is profound criticism of the argument from analogy, it implies two elements of social understanding that will be important in what follows. First, it is assumed that some form of inference is needed to access other minds:

c$_1$ – inference

The component of inference results from the assumption that mental states of the other person are hidden causes of perceivable behavior and thus need to be inferred.

Secondly, in order to infer mental states, it is asserted that a certain degree of similarity between individuals is needed:

c_2 – similarity
If social understanding partially draws on the application of self-related processes to another person, a specific degree of similarity between individuals is needed.

Again, these elements will be elaborated on throughout the course of this work. Let me now come back to Mill's argument and how it has been criticized. One important line of criticism is presented by Lipps, who contends that what is needed to understand another person is not information about one's own feelings, but about the other person's feelings. Instead of repeating, so to say, what I feel in a given situation, the task at hand is to gather information about what the other feels. Since this can be quite different from my own emotional state, no analogy is possible. In Lipps (1907, p. 708) words:

> D.h. ich soll nicht meine Trauer oder meinen Zorn, kurz mich, noch einmal denken, sondern ich soll etwas absolut anderes denken, nämlich an Stelle meiner und an Stelle meiner Trauer oder meines Zornes einen andern und den Zorn oder die Trauer eines andern, ich soll mich, das absolute Subjekt, vermöge dieses angeblichen Analogieschlusses vertauschen gegen etwas, das für mich Objekt und nur Objekt ist, ich soll diesen völlig neuen Gedanken eines Ich, das nicht ich, sondern von mir absolut verschieden ist, vollziehen.

Furthermore, Lipps argues that, since the way I perceive myself and others is fundamentally different, analogy cannot be the means by which I understand other humans. To grasp this approach, it is important to know that Lipps puts much emphasis on the role of expressions and claims that they exhibit a unique relation to emotions. Expressions are tightly related to emotions and therefore play an important role in knowing own emotions and that of others. Lipps claims that there is a fundamental discrepancy of how I gather knowledge of my own and the other's expressions and thus mental events. While the information about my own mind is rather drawn from proprioception, the source of information about the other person is gathered through visual perception (ibid.). Anger causes my face to change, the emotion is displayed in my facial features. However, I cannot *see* my own face while being angry, I at most 'feel' my facial expression change. When another person is angry, I can see her face change, but have no proprioceptive access to this alteration. This asymmetry of access leads to an asymmetry of how I conceive of my own and the others mind, leaving the conception of my own mind idle as a basis for inference. Lipps (1907) claims that I cannot know what the facial expression of the other means by analogy from my own case, because I have hardly seen my face in anger or sadness. Therefore, one would rather have to claim the opposite, i.e., that the visual perception of the other's face in anger or sadness provides the grounds from which I can infer the meaning of my own facial expressions. Analogical inference requires a common basis which simply does not exist due to the differences in access and thus is rejected as a mechanism to understand other minds.

Let me briefly describe one point of criticism of Lipps rejection of the argument from analogy. Stueber (2013) states that Lipps does not solve the problem of the argument of analogy, owing the reader an answer why Lipps own of account of empathy (which is now subordinate, but will be described in more detail in chapter 1.3.1.) does not encounter the same difficulties as analogical inference. Although Lipps names other mechanisms than analogical inference by which we are supposed to grasp foreign mental states, the discrepancy between the access to my own mind and to the other's mind is still given. According to Stueber (ibid., pp. 7–8),

> [t]he fundamental problem for Lipps's defense of empathy as primary method of knowing other minds consists in the fact that he still conceives of empathy within the context of a Cartesian conception of the mind tying our understanding of mental affairs and mental concepts essentially to the first person perspective.

The author is right insofar as Lipps indeed emphasizes the first person perspective as the only viewpoint one can ever have when it comes to experience mental events and emotions. However, it is questionable how this claim is a problem rather than a necessity. As I will show in greater detail in section 2.1.2., phenomenologists assert that a discrepancy or asymmetry of access to one's own and the mind of other's is a *constitutive* part of social understanding in virtue of yielding a necessary distinction between self and other. If it were not for asymmetry of access and thus boundedness to my own perspective, one would not be able to distinguish between one's own and other people's mental states. The problem with Lipps' account hence does not lie in the attempt to tie mental concepts to a first-person perspective, since it appears to be necessary to do so in order not to confuse self and other. We can thus list another component of social cognition:

c_3 – self-other distinction
Social cognition requires the distinction between self and other so not to confuse self- and other-related processes.

Another line of criticism is given by the phenomenologist Scheler. As Zahavi (2001) reviews, Scheler's critical points on the argument from analogy can be found in his work *Wesen und Formen der Sympathie*[6] and include an appreciation of Lipps' critique of the argument. Scheler first queries whether analogical inference can be attributed to any other being than a human adult. For Scheler, the process of inference is extraordinarily sophisticated. If one assumed that analogy were the only way creatures could socialize with their fellow beings, any animal – human or non-human – that is not able to draw analogies would be excluded. Thus, neither chimpanzees nor infants were able to gather knowledge about their social environment. Scheler (1912/1973, pp. 232–233) rejects this conclusion by giving examples of group behavior in great apes and infants that show social behavior at a very early age. As will be described in more detail elsewhere, both theory-theory and versions of simulation-theory have faced a similar critique in recent years (Gallagher, 2008). Although not always on the basis of analogy, the theories assume a quite sophisticated process which underlies interpersonal understanding, such as the formation of a theory or personal-level simulation (Goldman, 2006; Gopnik & Wellman, 1992). It is quite interesting to see how history keeps repeating itself; the controversy between phenomenology and mindreading approaches still revolves around the fundamental question of whether or not inferential processes are needed in order to make sense of one another.

Similar to Lipps, Scheler emphasizes a second point of criticism, namely the difference in access to one's own and the other's experience. There is, for example, no way to experience an emotion in exactly the same way that the other does: I do not have any proprioceptive information about the other's experience (Scheler, 1912/1973, p. 235). However, as Lipps has already pointed out, the argument from analogy only works if the experiential access of self and other is similar.

Closely related to this is Scheler's third critical observation that humans are able to interpret – at least to some extent – the behaviour of non-human animals. The mechanism by which we grasp, for example, that a dog's wiggling tail expresses excitement cannot be analogy since this kind of inference always presupposes psychological similarity, which is not given between the inferring human and the tail-wiggling dog. Thus, the argument from analogy fails to give an explanation of how an understanding of non-human creatures could possibly

[6] See section 2.1.1. for a clarification on Scheler's terminology and how he used many terms that all basically meant empathy.

occur. Lastly and most importantly, Scheler points to the argument's formal invalidity. If analogy is taken seriously, it never provides knowledge of the contents of the *other's mind* but only of *my own mind*; all I ever infer are my own mental states – but never foreign ones. Let me elaborate on this important point. In principle, analogical inference can be described in the following way: A certain emotion E triggers a facial expression F. When I observe F in another person, I infer that it was emotion E that led to F. An example would be the following: When I get angry (E), my face frowns (F). Thus, when I see a frowned face (E), I infer that the person must be angry (F). The crucial point now is that E represents *my own emotion* and *not the other's emotion* E_o. However, in order to understand the other's emotion, which may be different than mine, I would have to infer E_o. Only then would I know about her mind and not about my own. The argument from analogy presupposes a similarity (if not identity) of emotions that is not necessarily the case. In Scheler's (1912/1973, pp. 234–235) words:

> Denn logisch richtig [...] wäre ja der Analogieschluß nur dann, wenn er dahin lautete, daß, wenn gleiche Ausdrucksbewegungen da sind, wie ich sie vollziehe, auch *mein Ich hier noch einmal* vorhanden sei - *nicht aber ein fremdes und anderes Ich*. […] Der Analogieschluß könnte aber auf alle Fälle nur soweit zur Annahme fremder Iche führen, als diese meinem Ich gleich sind; niemals also zum Bestande fremder seelischer Individuen.

Therefore, by drawing analogies, one runs at risk of not understanding the other person, but merely what I would have felt or done if I were the other person.

Given this criticism, why should we keep inference and similarity as components of social understanding? I contend that the solution lies in disconnecting them from analogy. Inference can be seen in much more general terms and does not necessarily involve analogies. As will become more obvious, the notion of inference needed for social understanding merely refers to inferential mechanisms that can function without self-related input. Further, although a certain degree of similarity may be needed for social understanding, it is by no means a necessary condition for all cases of social cognition. These matters will be discussed in more detail in later chapters. For now, it shall be sufficient to note that the argument from analogy draws a picture of social cognition that would lead us – if drawing analogies were the only means to understand other people – to understand ourselves were we in the other's position, but not the other person herself.

1.2. Social Understanding as Theory Construction – The Case of Theory-Theory

What hese early accounts on how we understand each other that have just been described have in common is that they postulate the need for some kind of inference. This belief mostly stems from the conviction that minds are unobservable and that behavior only gives away so much about the other person's mental states. Turning towards a more modern approach to understanding each other, we will find that these ideas are still prevalent. In this chapter, I will introduce TT as such an approach, and as one that got the debate of social cognition going in philosophy of mind.

In order to understand how TT could arise as one of the main theories about social understanding, it will be helpful to describe some philosophical developments that took place when TT began to arise. This is also important because the very idea of TT grew out of general considerations about how the mind works. In the middle of the 20th century, new ideas about how the mind functions brought up new ideas of how intentionality could be characterized. Alan Turing introduced the thought that the human mind works like a computer and thus could be simulated by such a device. The mind-computer analogy gained importance as cognitive science began to grow into an interdisciplinary research field that took computer science, linguistics, and cybernetics as serious as neuroscience and psychology. As Metzinger (2010) claims, philosophy offered functionalism as a meta-theory that allowed the rise of new disciplines such as cognitive science or computational neuroscience. Functionalism tried to give an answer to questions that arose both in the debate about intentionality and the mind-body problem, namely how mental states (that have the feature to be about something, hence intentionality) can be conceived of as being part of the physical world (and thus, also the body). In answering this question, functionalist take the mind-computer analogy seriously and claim that mental states can be defined by the causal (or functional, hence functionalism) role they play for a system. Certain software in a computer can be realized by different hardware devices, as long as it fulfills a certain role. Likewise, mental states can – in principle – be realized by different devices. However, as Gardner (1985) notes, the most likely candidate for realizing mental states have soon been hypothesized to be neural states, thus transforming the mind-computer analogy to a brain-computer analogy and coining the internalist flavor of the debate.

These background developments are important for understanding how TT arose rather naturally from these convictions. In the search for an answer to the questions concerning intentionality, it has been a popular move to reconstruct the mind within a semantic model, thus applying the model of language to the mental realm. Two famous contributions came from the philosophers Wilfrid Sellars and David Lewis who followed this strategy. In this chapter, I will first describe how these two analytical thinkers tried to offer solutions to the problem of intentionality and mental representation in general, and in the process also provided possible solutions to the problem of other minds. Two important ideas that have changed both the theoretical and empirical landscape of research of social cognition shall be depicted:

(1) The idea that understanding others basically consists in forming theories about their mental states. In chapter 1.2.1., I will describe how Sellars' criticism of a Cartesian conception of the accessibility of inner mental states and his general account about those has given rise to one of the most prominent theoretical paradigms of social cognition in the 20th century: TT.
(2) Chapter 1.2.2. details the thought that the theory we draw on to infer mental states is a folk psychological, common-sense theory. I will be concerned with David Lewis' work on the functional status of folk psychology as the basis of theoretical inference.

After depicting the historical growth of TT, I will describe the further course of development of the theory.

(3) In chapter 1.2.4., I describe Premack and Woodruff's famous answer to the question: Does the chimpanzee have a theory of mind? Their work is important because they not only introduced the term of a 'theory of mind' (ToM), but also extended the scientific view to the non-human domain.
(4) In chapter 1.2.3., the so-called 'empiricist' version of TT will be described. This view assumes that the theory we have about other people and which enables us to understand them comes in the form of a 'theory of mind' (ToM) and works just like any other scientific theory.
(5) Chapter 1.2.5. deals with opponents of empiricist TT-versions who hold the belief that ToM is an innate capacity which is neurally realized by a ToM-module. This

view goes back on Gall's theory of phrenology and has considerable impact on further discussions about social understanding.

(6) Lastly, I present how developmental psychology further investigated the ability to form theories about other people's mental states. Chapter 1.2.6. describes how several psychologists carved the notion of ToM as a metarepresentational capacity whose development can be empirically examined and tested.

1.2.1. The Genius Jones Invents Theory-Theory

In this subchapter, my aim is to show that TT has its roots in the general philosophical idea that the mind is a linguistic device. Although Wilfrid Sellars was not particularly concerned with the problem(s) of other minds, his theory of the status of mental states has had important implications for the debate on social cognition. The following quotation contains several crucial points about the philosopher's views: "Sellars is not concerned with the mechanisms of thought – that is the bailiwick of the psychologist or even the neuroscientist – but with the nature of the concepts used in attributing thoughts to ourselves and others" (deVries, 2011, p. 22). First, deVries shows that *concepts* play a central role in Sellars' account, already hinting to the analogy he draws between language and the mind. Secondly, the attribution of concepts seems to be a strategy that is not only applied to others, but also to oneself. In a nutshell, Sellars is concerned with how mental episodes can be described and the presuppositions of our ability to talk about them. In his work *Empiricism and the Philosophy of Mind* (1956/1997), he elaborates on the view that one needs to develop a *theory* in order to be able to verbalize knowledge about inner episodes, i.e., about one's own and other's mental states.

I will now first describe how Sellars rejects the Cartesian view of knowledge about mental states as immediately given and exclusively accessible through introspection, which poses the problem of other minds (see chapter 1.1.1.). Sellars (ibid.) sets "the myth of Jones" against this "myth of the given" in order to deny the Cartesian view and to bring forth his own conviction that mental concepts are linguistic concepts and thus intersubjectively accessible. This will be described in the second part of the chapter, followed by the depiction of how this theory can be seen as the basis of one of the most prominent approaches to social cognition in the 20th century, TT.

To understand Sellars' motivation, some background information about his convictions will be useful. Sellars followed Kant in distinguishing thinking and sensing, hence rejecting the

Cartesian sensory-cognitive continuum. According to Descartes, both thoughts and sensory processing, as well as all other workings of the mind, basically have the same epistemological status. In this tradition, proponents of the sense-datum theory state that there are some cognitive states that are in direct contact with raw data of the external, sensory world, forming a foundation of knowledge. In this sense, what we perceive is immediately given (deVries, 2011). This also implies that our mental states, derived from this basis of knowledge, can only be accessed by the subject who has them. Inner episodes are thus thought to be not only absolutely private, but also infallible. As I have described in chapter 1.1.1. in more detail, the problem of other minds follows. How could it be possible to know the contents of other minds – or even to know that other people *have* minds – if they are exclusively accessible by one individual? This is a rough sketch of what Sellars (1956/1997) calls "the myth of the given" and which he attempts to replace by a more fruitful theory of thinking and knowledge.

Sellars basic assumption is that, in order to gather knowledge about the world, more than our senses are needed. In his opinion, language is the distinctive ability one needs to possess for both sensing and thinking. There cannot be knowledge without a preceding conceptual process. All knowledge requires concepts and classification, hence requiring language as its basis. As deVries (2011, p. 24) puts it: "In Sellars's view, thought is prior to language in the order of being, but language is prior to thought in the order of knowing." In his attempt to describe a version of the claim that thoughts are linguistic episodes, Sellars raises the question in which way these episodes can be linguistic without being uttered. In order to answer this question, he constructs the thought experiment of 'our Rylean ancestors' (Sellars, 1956/1997). These Ryleans are not only able to describe linguistic behavior, but are also in the possession of concepts to understand what other people are saying. What they lack is the ability to talk about inner mental episodes. After having described this people, Sellars (ibid., p. 92) asks

> What resources would have to be added to the Rylean language of these talking animals in order that they might come to recognize each other and themselves as animals that *think, observe,* and have *feelings* and *sensations*, as we use these terms?

At this point, it becomes obvious that in Sellars' view, language indeed precedes knowledge, since he assumes that our Rylean ancestors know the meaning of linguistic behavior before they know that its cause lies in their internal thoughts. The missing element, the capacity that our ancestors lack in order to be able to refer to the cause of their behavior, is the ability to build *theories*. This is when Jones enters the picture.

The life of our ancestors changes profoundly when a genius, Jones, appears. Sellars (1956/1997, p. 98) characterizes him as "[…] an unsung fore-runner of the movement in psychology, once revolutionary, now commonplace, known as Behaviorism". The behavioristic capacity Jones inhabits is described in more detail, in order to contrast this "methodological behaviorism" from "philosophical behaviorism" (ibid.). This kind of behaviorism neither rejects introspection nor does it make assumptions about the ontological status of mental concepts. Instead, it makes assumptions about how new concepts are construed. It furthermore holds that we introspect in terms of common-sense mentalistic concepts. The behaviorist acknowledges that a lot of knowledge is uttered in the form of mental concepts and realizes that even more knowledge can be gathered by putting hypotheses about behavior in common-sense mentalistic terms. The behaviorist will construct such concepts "from scratch" (ibid., p. 99), deriving them from the observation of behavior and then using them as heuristics. Sellars (ibid.) emphasizes that this cannot be the only way how concepts that are used to describe behavior are introduced; additionally, *theoretical* terms are needed.

In the following passage taken from *Empiricism and the Philosophy of Mind*, it becomes obvious how the introduction of Jones can be seen as the introduction of the basic idea of TT:

> Suppose, now, that in the attempt to account for the fact that his fellow men behave intelligently not only when their conduct is threaded on a string of overt verbal episodes – that is to say, as *we* would put it, when they "think out loud" – but also when no detectable verbal output is present, **Jones develops a *theory*** according to which overt utterances are but the culmination of a process which begins with certain inner episodes. *And let us suppose that his model for these episodes* which initiate the events which culminate in overt verbal behavior *is that of overt verbal behavior itself. In other words, using the language of the model, the theory is to the effect that overt verbal behavior is the culmination of a process which begins with inner speech.* (ibid., p. 103) [bold emphasis added]

The crucial point that has been picked up by proponents of TT is that, in order to describe mental states of other people, the development of a theory about them is needed. Let me hint to the points in this passage that are worth mentioning. Firstly, it is important to note that Sellars describes mental states as invisible in the sense that they are not directly perceivable by others, *if not uttered linguistically*. What makes Jones unique and ingenious is the fact that he not only finds a way to describe invisible inner states, but, first of all, to *realize* that his fellow men are still engaged in intelligent behavior even when there is no observable linguistic or behavioral output. Additionally, Sellars' conception of the relation between mental states and language once more becomes obvious. Talking is described as "thinking out loud" and further inner episodes are described as "episodes which initiate the events which culminate in overt verbal behavior." (Sellars, 1956/1997, p. 103) While mental states precede linguistic utterance, the possession of linguistic concepts precedes the ability to form a theory about mental states. Thus, the model for mental events is derived from the model of verbal behavior. Linguistic concepts are being used to describe non-observable, inner episodes.

This is not the end of the story of our Rylean ancestors and Jones. The genius now teaches his fellow human beings his strategy, which can also be applied to interpreting one's own behavior. Sellars uses the example of Tom and Dick: Tom is watching Dick and, making use of the newly acquired Jonesean skill, attributes the sentence "Dick is thinking that 'p'" to his fellow (ibid., p. 106). In the same manner, Tom is now in the position to attribute the sentence "I am thinking that 'p'" to himself, *given the same behavioral evidence* which caused him to attribute this sentence to Dick. Thus, "*what began as a language with purely theoretical use has gained a reporting role.*" (ibid., p. 107) In this sense, we can now see why Sellars describes mental concepts as genuinely intersubjective. The acquisition of the concepts we use to describe mental states (hence mental concepts), is profoundly based on linguistic abilities and linguistic concepts. We form these concepts – this is what Sellars calls methodological behaviorism – on the basis of the theories we build about the invisible intelligent behavior of other people and *only then* are able to use these concepts to describe our own inner mental events. In Sellars' (ibid., p. 107.) words:

> As I see it, this story helps us understand that concepts pertaining such inner episodes as thoughts are primarily and essentially *intersubjective,* as intersubjective as the concept of a positron, and that the reporting role of these concepts - the fact that each of us has a privileged access to his thoughts - constitutes a dimension of the use of these concepts which is *built on* and *presupposes* this intersubjective status. My myth has shown that the fact that language is essentially an *intersubjective* achievement, and is learned in intersubjective contexts – a fact rightly stressed in modern certain philosophers, e.g. Carnap, Wittgenstein – is compatible with the "privacy" of "inner episodes".

On the one hand, this implies a clear rejection of the sense-datum theory in the sense that all knowledge requires the 'medium' of language. There cannot be cognitive states that are in direct contact with the outside world and which thus form the basis of knowledge. Also, the view that mental states are accessible through introspection is denied. Not introspecting one's own mind informs us about our inner workings, but the formation of theories about the mental based on linguistic concepts. What remains, though, is the conviction that mental states are private and accessible to oneself in a privileged manner.

In conclusion, Sellars way of treating mentalistic concepts like theoretical concepts can be seen as an idea that gave rise to TT. What has been adopted by proponents of this theory is the view that in order to understand others, one needs to be able to build a theory, which in turn presupposes linguistic abilities. This is a very important feature of TT, since it has many implications for assumptions about what kind of creature is able to understand the behavior of its conspecifics. Strictly speaking, only beings that possess linguistic skills and terms were able to understand their fellow beings, thus excluding pre- or non-verbal creatures such as babies or non-human animals. TT has elicited much debate in empirical research about *how* this theory can be acquired (Gopnik & Wellman, 1992) and *when* it is acquired developmentally (Wimmer & Perner, 1983). But before this, the philosophical discussion focused on the status of the theory that was assumed to form the basis of attribution of mental states, folk psychology.

1.2.2. The Status of Folk Psychology

If understanding other people's behavior is achieved by forming a theory, what kind of theory lies at the heart of this ability? This question will turn out to be central not only for proponents of different versions of TT, but also for representatives of simulation-theory (ST). It was David Lewis who introduced the idea that the theoretical basis of understanding each other consists of *folk psychology, common-sense psychology* or *folk-science*[7]. In the following, I will describe three claims that are central to his account:

a) Theoretical terms are individuated by their functional role.
b) Common-sense psychology is the sum of platitudes from which we derive the meaning of mental states.
c) Learning about the functional role of mental states will inform us what they are.

[7] The term *folk psychology* has been used frequently in the 1980s to describe what David Lewis and others referred to as *common-sense psychology* or *folk-science*. I will use the three terms interchangeably.

While Sellars' focus lay on the analysis of concepts that are being used to describe inner episodes, Lewis' account makes ontological statements about the nature of mental states. The three claims result from Lewis' token identity theory of mental and neural states. In his paper *An Argument for the Identity Theory,* Lewis (1966) argues against a particular statement of psychophysical identity theorists. This statement consists of two components; the first is that the identification of neural and mental states is the same as the identification of theoretical concepts with the entities they describe (e.g., the identification of water with H_2O; Lewis, 1972). The second component is the claim that theoretical identifications merely serve parsimonious purposes and are "pieces of voluntary theorizing" (ibid., p. 249). In *Psychophysical and Theoretical Identifications*, Lewis (1972) argues that this is a false picture of the issue, since "[...] theoretical identifications *in general* are implied by the theories that make them possible – not posited independently" (ibid., p. 249). Lewis' view results from his functionalist conviction that theoretical terms are definable by the causal role they play in an empirical theory. This principle also holds for folk psychology:

> Applied to common-sense psychology – folk science rather than professional science, but a theory nonetheless – we get the hypothesis of my previous paper that a mental state *M* (say, an experience) is definable as the occupant of a certain causal role *R* - that is, as the state, of whatever sort, that is causally connected in specified ways to sensory stimuli, motor responses, and other mental states. (ibid., pp. 249–250)

Lewis gives the example of a detective who reconstructs a crime and in doing so introduces theoretical entities that have been unknown to the listener of the detective's story until she hears the story. The detective starts to theorize about three suspects that could have committed the crime and assigns "T-terms" (theoretical terms)[8] (ibid., p. 250) to them by naming the suspects X, Y, and Z. While listening to the reconstruction of the crime, we learn some facts about X, Y, and Z and thus come to know what they refer to. In telling a story, the detective attributes certain *roles* to the suspects. That X, Y, and Z *realize* a certain role would still be true if the detective began to use the suspects proper names (Plum, Peacock, and Mustard), but Lewis states that they *uniquely realize* the functional role they inhabit, which means that the story would not be true with a different triple of names. He concludes that:

[8] Lewis contrasts "T-terms" with "O-terms", 'o' standing for 'old', 'other', or 'original' Lewis (1972, p. 250). He emphasizes that 'o' does not refer to 'observational', criticizing Sellars (1956/1997) who set up this distinction.

> They were introduced by an **implicit functional definition**, being reserved to name the occupants of three roles. When we find out who are the occupants of the three roles, we find out who are *X, Y* and *Z.* Here is our theoretical identification. (Lewis, 1972, p. 251) [emphasis added, LQ]

This is how Lewis argues for the first claim a) that theoretical terms are individuated by their functional role. The same principle applies to concepts of mental states. They gain their meaning by being ingrained in a certain kind of theory, common-sense or folk psychology. Mental terms are used in a particular way, hence inhabiting a certain functional role. Folk psychology is the theory we use to explain people's behavior and from which terms of mental states are derived:

> Think of common-sense psychology as a term-introducing scientific theory, though one invented long before there was any such institution as professional science. Collect all platitudes you can think of regarding the causal relations of mental states, sensory stimuli, and motor responses. Perhaps we can think of them as having the form: When someone is in so-and-so combination of mental states and receives sensory stimuli of so-and-so kind, he tends with so-and-so probability to be caused thereby to go into so-and-so mental states and produce so-and-so motor responses. (ibid., p. 256)

In this quotation lies the second claim b) saying that folk psychology is the sum of platitudes from which the meaning of mental states is derived. In a next step, Lewis claims that learning about the functional role of mental states will inform us about what they are (claim c)). This is achieved in the same fashion in which we learned about the suspects in the detective's story or about the fact that water is H_2O (Lewis, 1972), namely theoretical identification. Stephen Stich (1986), who proposes a kind of eliminativism of folk psychology in his book *From folk psychology to cognitive sciences*, notes that this claim has several implications.

Firstly, this account of TT is committed to the truth of folk psychology: "If the folk theory [...] turns out to be false, then the mental state predicates the theory implicitly defines will be true of nothing" (ibid., p. 21). Secondly, the theory is ontologically non-committal in that it leaves open the kind of entity that fulfills the causal role in question. The most likely kind of entity, according to theory-theorists, are neural states. This speaks for token identity theory rather than type identity theory, since the same mental state can be realized in different individuals by different brain states, as long as this neural setting still fills the same *causal role* (cf. ibid.).

Again, what we find here is the assumption that understanding others involves theoretical inference. Lewis' theory of the status of folk psychology adds an important component to this view, namely the claim that the theoretical terms that are needed are derived from a set of folk psychological platitudes. The question about the nature of the theories we construct in order to understand others has been a central point of discussion ever since.

1.2.3. Non-Human Theory of Mind

Another important part of the discussion was elicited by the work of Premack and Woodruff (1978) who asked whether the ability to theoretically infer the mental states of others is confined to the human species. Their work not only further refined the concept of a theory, but also introduced the term 'theory of mind', which is ever since widely used to refer to certain social cognitive abilities. Furthermore, their definition of ToM became broadly accepted and elicited a great amount of empirical research on the topic in developmental psychology, psychiatry and neuroscience.

Premack and Woodruff (ibid., p. 515) wanted to test whether chimpanzees possess ToM, i.e., the ability to infer the mental states of another individual, or, in their words: "In saying that an individual has a theory of mind, we mean that the individual imputes mental states to himself and to others (either conspecifics or to other species as well)." They hypothesized that ToM is the means by which humans understand each other's behavior:

> In assuming that other individuals *want, think, believe,* and the like, one infers states that are not directly observable and one uses these states anticipatorily, to predict the behavior of others as well as one's own. These inferences, which amount to a theory of mind, are, to our knowledge, universal in human adults. (ibid., p. 525)

Premack and Woodruff qualify their choice of terms by claiming that it is appropriate to call this ability a theory for two reasons, the first being that mental states are hidden, the second being that inferences are used to predict behavior. In order to test whether chimpanzees possess this strategy, they showed Sarah, a young female ape, different videos that all display a human who is unable to access a piece of food. Before the human succeeds in grabbing the banana, the video is put on hold and Sarah is presented with four different photos that all show possible outcomes of the situation. Unambiguously, the chimpanzee chose the photo that depicts the solution to the problem of inaccessible food. The most likely explanation of this is, according to the authors, that Sarah is capable of two things, both proving that she possesses ToM. Firstly, she must have attributed an *intention* or *purpose* to the actor.

Secondly, she must have known about the actor's *knowledge* or *belief*.[9] The authors conclude by questioning the development of ToM. In their view, the ability is not acquired during a teaching process, but the "[…] acquisition is more reminiscent of that of walking or speech. […] All this is to say that theory building of this kind is natural in man." (Premack & Woodruff, 1978, p. 515) This view further promotes the assumption that theoretical inference is the main ability by which individuals understand each other. By attributing ToM to non-human animals and thus non-linguistic creatures, an interesting alteration in how to conceive of ToM takes place. The ability cannot longer be seen as a fully linguistic one which requires high-level, symbol-like conceptual mental states. Much rather, the concept of a theory is stretched, which triggers the question of the nature of the theory and its implementation anew.

Against this background, Premack and Woodruff furthermore raise the question whether children and children with mental disorders possess ToM. It is this issue that has been picked up by psychologists and neuroscientists in the 1980s, such as Josef Perner and and Heinz Wimmer (1983), who developed the "false-belief task", which has become one of the most important empirical paradigms. Uta Frith, Simon Baron-Cohen and Alan Leslie (1985) tested whether children with autism spectrum disorder possess a theory of mind, famously hinting to the fact that one major impairment in autism lies in difficulties with social cognitive skills as will be described in chapter 1.2.6.

1.2.4. Empiricist Theory-Theory

One famous attempt to answer the question of the kind of theory that is exhibited in TT will be described in the following. It can be found in Gopnik and Wellman's (1992) paper *Why the child's theory of mind really is a theory.* The authors aim to reach two goals in their work: (1) spelling out the nature of theoretical entities that underlie folk psychology, and (2) describing the development of the child's theory of mind in analogy to scientific theories. Goldman (1989) criticizes TT for not offering an accurate and plausible description of the

[9] The authors also consider the possibility that Sarah's right answers have been due to associationism, i.e., familiarity with the situation, or empathy, i.e., the ability to put herself in the actor's place. Although they claim that the three possibilities are not exclusive, they also claim that associationism and empathy are more likely to appear in familiar situations or situations in which the attribution of motivation is sufficient. But for inferring the solution of a novel problem – like the one Sarah was presented with in this study – and for inferring the beliefs of the actor, a more sophisticated skill, namely theory of mind, is needed (Premack & Woodruff, 1978).

theoretical entities that amount to the folk psychological set of laws used in mental state attribution. Gopnik and Wellman aim to fill this gap by claiming that these entities have the same properties and features as any other theoretical construct used in scientific theories. Theoretical constructs *postulate* invisible entities, i.e., they are *abstract* descriptions of phenomena. Furthermore, theories are explanatory, predictive and prone to error. The last point leads to the necessity for theory change, which has its own dynamics: After the accumulation of counter-evidence for the prevalently held theory and the rejection of this evidence as mere noise or coincidence, ad hoc explanations serve to try to further keep up the original theory and ignore counter-evidence. Finally, the theory is revised and modified in light of new evidence. According to Gopnik and Wellman (1992, p. 149), this is exactly what can be found in children:

> Children should ignore certain kinds of counter-evidence initially, then account for them by auxiliary hypotheses, then use the new theoretical idea in limited contexts, and only finally reorganize their knowledge so that new theoretical entities play a central role.

Going into more detail, the authors describe the development and emergence of ToM between the age of 2.5 and 5. Infants that are not older than 2 years are already thought to possess early notions of mental states. The authors claim that this is suggested by the findings of Trevarthen (1979) and Meltzoff (1977), showing that young infants engage in imitation and joint attention. All of these findings are interpreted as pointing to crude and early possession of mental concepts. Thus, while 2-year olds only possess an incorrect early version of ToM that is better described as non-representational,[10] 3-year olds are at the stage of making ad hoc auxiliary hypotheses and expand their cognitive mental terms. They do acknowledge the existence of the mind and representational states, yet they cannot make proper sense of them. Although children at that age start grasping notions of 'non-real' mental states like dreaming and pretending, their beliefs are still held in a non-representational manner. Only between the age of 4 and 5 do children acquire a "representational model of the mind [and] the child's view of the mind becomes fully intentional" (Gopnik & Wellman, 1992, p. 153) in terms of propositional attitudes and

[10] It is claimed that children at that age are able to grasp perception and desire as "two basic categories of explanatory entities in folk psychology" (Gopnik & Wellman, 1992, p. 149). In order to make sense of perception and desires, the child does not need to conceive of them as complex representational states, but can depict them as simple causal links between mind and world. Note furthermore that a non-representational view also excludes the possibility of misrepresentation.

propositional content. After reorganizing their theory, children start realizing that thinking determines action and behavior. This is called a representational model of the mind because representations are thought to guide psychological functions.

In summary, this version of TT has been called empiricist TT because ToM is seen as a proper scientific theory that is acquired in an empirical manner and possesses the same properties as any empirically built theory. Although the empirical process thusly lies at the heart of ToM, Gopnik and Wellman do not deny that there could also be some innate form of ToM that serves as a basis for further theory construction. Whether or not ToM is an inborn ability or whether it is acquired during development was one of the most pressing questions at the time. In the next chapter, I will introduce a view that is much more radical about the idea that ToM is innate.

1.2.5. Nativist Theory-Theory

So-called nativist versions of TT mostly build their claims of a theory that concerns the architecture of the brain and mind, viz., *modularity of the mind*. Let me elaborate on this more general view of the mind before I show how it relates to ToM. The basic idea behind a modular view is that the mind is organized in distinct units that serve different functions. Historically, modularity goes back to the theory of phrenology. Introduced by Franz Joseph Gall and colleagues in the 18th century, the basic idea is that the brain is the organ of the mind, composed of different areas in which psychological processes can be located. Although most of the details of phrenology have been rejected (e.g., that brain areas are not connected) as neuroscience gathered deeper insights and more sophisticated methods, the theory reflects important intuitions and insights that have driven research until today. Firstly, in locating the mind within the brain, Gall opposed the then prevalent Cartesian view, according to which the body (and thus, the brain) and the mind are two distinct entities. Although Descartes intuited that there must be some relation between brain and mind by naming the pineal gland as the location in the brain where mind and body interact, Gall's approach goes many steps further. He does not claim any ontological separation between psychological and bodily processes, but a direct relation between the brain and psychological traits (cf. Adolphs, 2010a, p. 757). This points to another important element of the theory; viz., the intuition of specialization. In principle, the conviction that brain areas are specialized still holds today. This view has of course been refined, for example by showing that although specialized, brain areas still interact or underlie several rather than only one psychological process.

However, while phrenology is a theory that is *primarily* about the organization of *the brain,* the theory under recent scrutiny – the modularity of the mind – *primarily* makes assertions about *the mind*. Still, phrenology depicts the brain as *the organ of the mind*, and ascribes particular psychological functions to particular parts of the brain.

Jerry Fodor (1983), whose book *The modularity of the mind* can be seen as the first modern milestone of modular views of the mind, picks up this idea. In claiming several features of the modular mind, Fodor fueled a lively debate not only about the architecture of the mind, but also of the brain and how the two are connected. Fodor describes different modules for different mental functions. Innateness is a crucial property of these modules, meaning that the abilities they inhabit "develop according to specific, endogenously determined patterns under the impact of environmental releasers" (ibid., p. 100). One example of such a module is the ToM-module, which enables humans to make sense of others. There are several other theories of ToM that draw upon a modular view of the mind. An example is Leslie's (1994) view, in which he claims that there is a theory of mind module, which is specialized for the processing of meta-representations. Another example comes from Simon Baron-Cohen (1997), who claims that there are four different modules that underlie our capacity to predict and interpret the behavior of others: the eye-direction detector, the intentionality detector, the shared-attention mechanism, and the theory of mind mechanism.

Furthermore, modules are described as *domain specific* (i.e., they specifically inhabit a particular function), and *informationally encapsulated* and *inaccessible* (i.e., they cannot be guided by information from other modules). Drawing on phrenology, Fodor further claims that each module is *neurally localizable*, which means that modules have their own fixed neural architecture. Related to these ideas, modules are moreover *functionally dissociable*. Functional dissociation implies that a module can be impaired, but leaves the function of other modules intact.

The view just described influenced the debate about TT and social understanding in more general in several ways. First, while Gopnik and Wellman emphasized the growing sophistication of ToM by gathering scientific evidence, Fodor states that ToM as such does not undergo any profound change. The child, according to him, in principle has the same ToM-ability as an adult. The only thing that changes during development is her ability to exploit this innate capacity (Fodor, 1992). Further, one of the most important hypotheses that drives research about social cognition in the social neurosciences until today is strongly reminiscent of a modular view of the mind, the *social brain hypothesis* (SBH; see chapter 3.2.1.). SBH states that there are parts of the brain that function for social cognition; the sum of these brain regions is then referred to as the social brain. I shall not go into detail about SBH at this point, but just emphasize once more that the modular view of the mind had a profound impact on how researchers and philosophers thought about the mind and also social understanding.

1.2.6. Developmental Psychology and Theory of Mind

After having presented how the concept of ToM was introduced into the debate on mental state attribution and how it was thought to be acquired, I will now focus on how the conceptual landscape developed further. On the one hand, ToM has been further specified as a *metarepresentational* skill, thus implicitly confirming the worry that ToM may be a high-level ability that is only available to a specific group of individuals. On the other hand, the term 'mentalizing' has been thrown into the discussion as an alternative term for the ability to attribute mental states to other people.

Before presenting their experimental design that has fueled much discussion in philosophy, psychology and neuroscience, Heinz Wimmer and Josef Perner (1983) clarify their conception of ToM, referring not only to Premack and Woodruff, but also certain philosophers. In their view, the capacity to understand other people's mental states and propositional attitudes is *metarepresentational*, meaning that one does not only need to be able to represent the propositional attitude in question, but also be able to *represent* the fact that she *represents* this certain belief, desire, etc. Wimmer and Perner (ibid., p. 104) draw upon Pylyshyn's notion of metarepresentation:

> In Pylyshyn's (1978) explication this means that somebody who has a theory of mind does not only have a representation about a state of affairs (x) and stands in certain relationships to these representations (e.g., wanting x, believing x, etc.) but also represents these relationships explicitly. Pylyshyn refers to this ability as an ability for 'metarepresentation'.

One way to test whether a being has metarepresentational abilities is to test whether she or he is capable of deception. As Wimmer and Perner show in more detail, Premack and Woodruff taught one of their chimpanzees to deceive a competitor. Additionally, before the false-belief task has been conducted for the first time, there have been experiments which showed that very young children are able to verbalize and explicitly represent the relation in which they themselves and others stand to the propositional content in question (Wimmer & Perner, 1983). If the belief or desire that the other person holds is different from the own belief or desire, things become even more complicated. The capacity of metarepresenting a propositional attitude that is *different* from the one I myself am holding is exactly what is being tested in the false-belief task. I will shortly describe the main experiment.

In order to test at which age ToM as a metarepresentational ability arises, normally developed children of three age groups were presented with the following scenario. 3-4-year olds, 4-6-year olds and 6-9-year olds watched a play of puppets: A boy, Maxi, and his mother are in the kitchen. The mother bought chocolate that she puts in one of three boxes, the blue one. Maxi, who loves chocolate, plans on taking some chocolate later on. He leaves for the playground, while his mother starts to bake a cake. She uses some of the chocolate, but when she puts it back, she puts it in the green box and leaves the kitchen. After some time, Maxi returns from the playground and hungrily decides to eat chocolate. At this point, the children are asked three questions. The two control questions – that have been answered correctly by all children in all age groups – make sure that they remember where Maxi put the chocolate ("Do you remember where Maxi put the chocolate in the beginning?" *Memory Question*) and test if they know where the chocolate really is ("Where is the chocolate really?" *Reality Question*). The crucial question – described as the *Belief Question* – asks: "Where will Maxi say the chocolate is?" If the children point to the blue box – the one where Maxi *must (falsely) think* the chocolate is, since he hasn't seen his mother replacing it – they pass the test, because they were able to attribute a false belief of the location of the chocolate to Maxi. All children between the age of 6 and 9 passed the test, while almost 60% of the children between the age of 4 and 6 answered correctly. None of the children between 3 and 4 succeeded in attributing the false belief to Maxi, all of them pointed to the green box, the one in which the chocolate really was. The authors conclude that the ability to understand another person's belief, even if it is a wrong one and different from what oneself believes, seems to rely of the emergence of a newly acquired cognitive capacity around the age of 4:

> In summary it seems, therefore, that the emergence of children's ability to understand another person's beliefs and how this person will react on the basis of these beliefs and their understanding of deception is not a mere side effect of an increase in memory and central processing capacity. Rather, a novel cognitive skill seems to emerge within the period of 4 to 6 years. Children acquire the ability to represent wrong beliefs and to construct a deceitful or truthful utterance relative to a person's wrong beliefs. (ibid., p. 126)

Simon Baron-Cohen, Uta Frith and Alan Leslie (1985) can be seen as early adopters of the false-belief task. Their interest focused on testing whether children with autism spectrum disorder (ASD) possess ToM. Several assumptions led the scientists to the conviction that the false-belief task would be an appropriate tool to test autistic children and their social abilities. Firstly, social impairments are not related to IQ, reinforcing the assumption that there must be something specific about them. This becomes obvious when comparing children with ASD and children with Down's syndrome. The latter usually show significantly lower IQs, but no social impairments. Thus, the degree of mental impairment is no predictor for social (dis)abilities. Furthermore, the authors frame their hypotheses in the metarepresentational model of ToM which was described by Leslie and which – equally to Wimmer's and Perner's theorizing – relies on the notion of metarepresentation or second-order representations set out by Pylyshyn (1978). According to Leslie (1987), the ability of metarepresentation transforms into two additional abilities, namely ToM and pretend play. Thus, if the first (i.e., the ability to metarepresent) is lacking, neither of the latter abilities could arise. This relates to earlier studies showing that autistic children engage less in pretend play or show impoverished skills in doing so than their typically developing peers (Baron-Cohen, Leslie & Frith, 1985).

The central hypothesis of the authors is that – since an explanation of the fact that the specific impairments in the social realm is still missing – the proof of a lack of metarepresentational abilities in autistic children could fill this explanatory gap. Since the possession of ToM is seen as a metarepresentational ability, the lack of ToM would also proof a lack of second-order representation. Moreover, the false-belief task is seen, at the time, as the strongest empirical tool to test whether children are in the possession of ToM or not. In their version of the task, Baron-Cohen and colleagues (1985) put three different groups of subject to the test: one group of high functioning autistic children, one of children with Down's syndrome and one group of neurotypical children. All groups tested equal in mental age and IQ and were presented with a similar scenario than Wimmer's and Perner's false-belief task. This time, Sally and Anne are playing and while Sally leaves the room, Anne replaces a marble to a different place. After being asked the Reality and Memory Question, which were successfully answered by all participating children, the Belief Question followed: "Where will Sally look for her marble?" In order to pass the test, the participants had to predict the doll's behavior on the basis of her false belief. The results show that while 85% of children with Down's syndrome and 86% of normally developed children answered correctly, 80%

of autistic children failed the test. The authors conclude that children belonging to the latter group were not able to dissociate their own knowledge of where the marble really is and the doll's wrong belief of the marble's location (Baron-Cohen et al., 1985). The hypotheses about a lack of ToM as a metarepresentational capacity and its relation to other symptoms of autism that hint to social deficits such as a lack of pretend play is thus seen as confirmed. Let me now point to some important implications of the interpretation of these findings. Wimmer and Perner are quite specific about their explanatory scope. The empirical paradigm they conducted aims to test the ability to represent wrong beliefs of others and to determine the point in development at which this ability arises. The successful representation of the *difference* between the correctness of one's own belief and the falsity of the other's belief is seen as a complicated metarepresentational process whose presence or lack is tested in the false-belief task (Wimmer & Perner, 1983). This assumption is shared by Baron-Cohen, Frith and Leslie. However, while Wimmer and Perner depict ToM as a general cognitive function, Baron-Cohen and his colleagues relate the ability to *social* cognitive functions. Their aim in applying the false-belief task to autistic children is to find out whether their *social* impairments can be seen as standing in a causal relationship to a lack of metarepresentational abilities: "Thus we have demonstrated a cognitive deficit that is largely independent of general intellectual level and has the potential to explain both lack of pretend play and social impairment by virtue of a circumscribed cognitive failure." (Baron-Cohen, Leslie & Frith, 1985, p. 44)

This hints to three further implications: Firstly, Baron-Cohen and others describe social abilities as something that is *dissociated* and different from other, more general cognitive features such as IQ. This is a first step towards the assumption that there is something special about the abilities in question. Secondly, both groups of authors frame ToM as a general metarepresentational ability, which is very different from how ToM is seen nowadays, namely as a *genuinely social* function. Baron-Cohen, Frith and Leslie start to think of it as an ability that is at least related to more *specific* social abilities in important ways. Thus, taking the first and second implication together, this line of thinking can be seen as a precursor to the conviction that drives much contemporary debate and depicts social cognition as crucially different from general cognition. As will be detailed in later chapters, this assumption is crucial for justifying a genuine research field for the phenomenon. At this point, we can thus formulate a first desideratum for a theory of social cognition:

d_1 – specificity
Specificity postulates the need for the assumption that there are properties of social cognition differentiating it from general cognition.

Of course, there is much debate about whether or not social cognition is different from more general cognitive processing. This is why I deem it of high importance for a theoretical take on social understanding to consider the possibility of specificity. As for now, suffice it so say that the debate on social cognition started to grow into a proper field of research. The claim that there is something special about the phenomenon that distinguished it from general cognition is vital to this development, as will be detailed in chapter 3.3.1.

Lastly, while the authors do not see ToM as a social ability *per se,* it is certainly seen as a necessary ability to possess in order to develop proper social capacities. This implies that metarepresentational abilities are crucial for and may even be a presupposition for developing social abilities.

I briefly want to point to a term that has been introduced by Uta Frith and colleagues (1991, p. 434) in order to account for ToM abilities as the ability to attribute mental states to others, namely *mentalizing*:

> This is the development of the 'theory of mind', or 'mentalizing' - our ability to predict and explain the behaviour of other humans in terms of their mental states. Our ability to mentalize is revealed in our use and understanding of such words as 'believe', 'know', 'wish', 'desire', 'intend', and 'pretend'. A central feature of our proposal is that children with autism lack this ability.

Mentalizing is furthermore seen as the ability that underlies fluid social interaction in everyday life. Here one can see a clear step towards the assumption that ToM (or, mentalizing) is a major feature of social cognition. Mentalizing in this formulation is still seen as a metarepresentational ability, which reinforces the assumption that major parts of social cognition rely upon high-order representational skills. However, although Frith yields a definition of mentalizing, its exact relation to ToM remains unclear. While the quotation above suggests that the terms are interchangeable, at a different point of the paper, mentalizing is described as the ability that a child needs to possess *in order to develop* a ToM (ibid., p. 437), hence depicting the first as a precondition for the latter.

In sum, although the nature of TT remains unclear until today, although there is much criticism on TT (see chapter 1.3.2.) and although it is uncertain whether or not ToM is a metarepresentational capacity, we can hold on to the following. One component of social cognition is the ability to construct theories about the mental states of other people:

c_4 – construction
Construction refers to the ability to consciously construct theories about the mental states of other people.

As will turn out at later stages in this thesis, we can make this statement without committing to versions of TT. Theoretical inference here refers to the high-level ability to consciously construct a theory about the reasons of another person's behavior. In contrast to the versions of TT that I have just presented, I do not claim that theory construction is the *only* or even the *main* mechanism of understanding others. In the following, I will keep on describing critical points about TT that make my denial of this claim more obvious.

1.3. The Historical Growth of Simulation-Theory

As described in the previous section, TT mainly offered a view on the *kind of knowledge* we retrieve when we are trying to understand each other. Proponents of ST not only offer an alternative proposal for this, but also start to focus on the mechanisms and processes that are being used in applying whatever kind of knowledge to the other person.

The intuition behind the idea of ST is that we use (implicit or explicit) knowledge about *ourselves* and *project* it onto the other person in order to understand, predict, or interpret her behavior. In this section, I will describe the historical growth of ST and point out important differences to TT. In doing so, I proceed as follows:

(1) I start with a very early version of ST that did not carry the name, but certainly entails central claims of later simulation theorists. In chapter 1.3.1., Lipps' conception of empathy is described, who claims that understanding others is based upon the urge to internally imitate their emotions.

(2) Before I go into more detail about the positive claims of ST, I first present the critical points that simulation theorists voiced about TT in chapter 1.3.2. Specifically, I scrutinize their criticism about folk psychological knowledge as nomological rules.

(3) In chapter 1.3.3., Gordon's account of ST is depicted. What is interesting about his theory is that he describes simulation as a process-driven capacity that can also be 'taken offline' and operate on a sub-personal level.

(4) Philosopher Alvin Goldman is probably one of the most famous proponents of ST. In chapter 1.3.4., I describe his first version of how simulation is supposed to work. Goldman emphasizes that there needs to be an isomorphism between the simulator

and the person who is being simulated and also introduces the idea that understanding others may imply both simulation and theoretical inference.

(5) After the discovery of mirror neurons in the early 1990s, simulation theorists thought to have found the neural basis for the mechanism. In an influential paper, Goldman and Gallese present their theory of mirror neurons and the simulation theory of mindreading. This will be described in chapter 1.3.5.

(6) Lastly, I detail Gallese's more detailed account of how the mirror neuron system enables social understanding via simulation. Importantly, his account is an attempt to couch simulation into neuroscientific terms and into a framework that describes social understanding at multiple levels of description.

1.3.1. Lipps' Conception of Empathy

How exactly does Lipps describe empathy? In his *Leitfaden der Psychologie,* Lipps (1903) claims that there are three areas of insight; one can gain knowledge, firstly, about objects and things through sense perception, secondly about oneself through inner perception and retrospection, and, thirdly, about other individuals. In the latter case, empathy is the primary way to gather knowledge (cf. Lipps, 1903, p. 191). The use of the word 'Einfühlung' or empathy[11] stems from romantic German novelists such as Herder and Novalis who used it in a loose manner to describe the general ability to 'feel into' nature (Stueber, 2013). Lipps' conception of empathy is equally broad, referring not only to the capacity to understand other minds, but also to a means to aesthetically grasp architecture, arts or nature (Metzinger, 2009, p. 171). Although phenomenologists have criticized his positive account of empathy, they nevertheless adopted the term that is still being used in current research to describe the understanding of emotions of other persons.

As already implicit in his statement that there is an asymmetry of access to self and other, Lipps argues that the empathic understanding of others has its own perceptual quality, and that "our knowledge of others is a modality of knowledge *sui generis*, something as irreducible and original as our perceptual experience of objects or our memory of our past experiences" (Zahavi, 2010, p. 288). Attributing a specific perceptual quality to empathy is a move which is important for the broader debate on social cognition for several reasons.

[11] The English term empathy has been introduced by Edward Titchener in the year 1909 (Stueber, 2013).

First, recall that specificity has been named as one desideratum for any theory of social cognition. Lipps' observation that there is something experientially specific about social cognition underlines the relevance of this theoretical requirement, since it assigns specificity at the phenomenological level of description. What is more, the unique experiential quality of social encounters is a point that has been emphasized in the phenomenological tradition. Whilst I will amply criticize the phenomenological perspective (see chapter 2.2.3.), I still consider it important to pay close attention to the experiential quality of social cognition. For example, it may turn out that the phenomenology of social cognition is indeed different from general cognition, thus providing a criterion that speaks for the specificity of the phenomenon. Thus, let me list another component of social cognition that shall not be missed in a theoretical account of the phenomenon:

c_5 – experiential quality
Experiential quality refers to the assumption that social encounters come with different phenomenal signatures that serve a specific function.

In an attempt to describe the phenomenal nature of empathy, Lipps (1903, p. 192) states that there is no *immediate* visual or auditory perception of the mental states of others, but that we simply experience them *within ourselves* and *through ourselves*: "Wir sehen nicht, noch hören wir das Fühlen, Vorstellen, Wollen eines Anderen, und das Individuum, das vorstellt, fühlt usw. Sondern wir erleben dergleichen einzig in uns. Aus den Zügen der eigenen Persönlichkeit müssen wir also die fremde weben.“

The crucial question then is how the basis of empathy that lies within ourselves is established. Lipps answers this question by stating that empathy has two components which ultimately lead to the comprehension of the other's expressions. The first component is the urge to imitate, the second is the urge to express psychological experiences in a certain manner (Lipps, 1903). The former depicts the core of Lipps account and has also been called "instinct of empathy" (1907, p. 713); instinctively, one reproduces the observed movements, expressions and gestures. This automatic tendency of reproduction elicits the associated emotion in oneself, which then is projected onto the other. In other words, inner imitation based on one's dispositions for motor mimicry and the ability to project the elicited emotions onto the other enable humans to attribute meaning to the observed expressions, gestures or movements of others. As we shall see, this is one component of social cognition that researchers from all fields agree is crucial; phenomenologists, psychologists, philosophers and social neuroscientists at some point all come to an understanding that when we socially interact with others, or even merely observe them, some kind of 'mirroring' takes place. This means that part of the other's state (e.g., her bodily or emotional state) is reproduced and that this replication is crucial for the process of understanding or empathizing with her. We may thus introduce the next component of social cognition:

c_6 – replication
Replication describes social cognitive processes that involve reproducing the state of the other person in order to gather information about her.

This aspect will come up again throughout this thesis in different conceptual clothing. Lipps called it the instinct of empathy, later it was called mirroring or simulation, some may call it automatic imitation or mimicry. I chose *replication* as an umbrella term for these phenomena which entail that part of the other's state is replicated.

The 'instinct of empathy' relates to another important aspect of the phenomenon, which has been picked up by later proponents of ST, viz., the role of the body for empathy. Lipps account entails the claim that while the body of the other serves as the source of information and builds the object of observation, one's own body serves as the location in which the other's mental states are reproduced via imitation and/or motor mimicry.[12] Although we may not find any assumptions about the necessity or sufficiency of the role the body is playing for empathy, and whilst this element will be described in much greater detail in later parts of this work, Lipps still put enough emphasis on bodily processes. These processes shall be considered as an important component of social cognition:

c_7 – embodiment
Embodiment refers to the assumption that the body plays a crucial role for most social cognitive processes.

Of course, the term 'embodiment' was not part of the conceptual repertoire of Lipps and would not turn into a commonly used notion until much later after his death. However, the

[12] A similar view can be found in Vittorio Gallese's (2005) account of embodied simulation (for a closer description, see chapter 1.3.6.). Gallese as well states that observing the other elicits an automatic and embodied process of imitation, which reproduces the mental state and projects it back onto the other individual. Also, both authors emphasize the automaticity of imitation, without effort or planning, humans are urged to internally mimic the mental states expressed in the behaviour of another person, which builds the basis of interpersonal understanding.

term still captures the idea that bodily process play an important role in understanding other people and also points to the larger movement of embodied cognition.

To sum up, Theodor Lipps philosophical account of empathy as an epistemic tool to grasp other minds has influenced the debate of social cognition sustainably. Not only has he elicited the discussion of the right mechanism that underlies understanding others, but he also introduced empathy as a term that has been used widely until the cognitive turn. His idea of an inner urge to imitate is still meaningful for ongoing debates and views of social cognition.

1.3.2. Simulation-Theory vs. Theory-Theory

Before I describe two versions of ST in more detail, let me first show how two pioneers of early ST – Gordon and Goldman – have criticized TT and thus deemed a new view on understanding the mind and others necessary.

As will be clear soon, both authors endorse that the prediction and interpretation of the behavior of others happens on the basis of a simulation process rather than the application of a theory based on generalized, folk psychological laws. It is exactly this point of TT that is being criticized as implausible by proponents of ST. Gordon's criticism focuses on the fact that an underlying set of *universally valid* laws could only explain *successful* prediction of behavior – one's own as well as the other's. He argues that we do make wrong predictions about how we will behave in a certain situation and that this speaks against the assumption that we do so by drawing on a reliable nomological basis. These failures of prediction, in his view, speak for practical simulation. However, while it seems plausible that the application of rules brings forth only correct predictions, this does not actually follow. It could well be that folk psychological rules are applied *inappropriately* or that the wrong laws are being used to explain a certain behavior. There is no reason why TT is committed to the statement that only correct predictions or inferences are drawn.

Gordon goes on to criticize that TT is confined to cases and situations in which generalized laws actually *hold*, i.e., typical situations, but cannot be related to atypical or deviant ones. Thus, theoretical inference on the basis of folk psychological laws has a very limited explanatory scope when it comes to interpreting behavior in general (Gordon, 1986). Goldman (1989), who – other than Gordon – rejects the functionalist background assumptions of TT, has yet another point for claiming that TT operates on implausible presumptions. According to him, no accurate or persuasive description of folk psychological laws has been given, although a functionalist version of TT – such as Lewis' – requires not only that folk psychology is *true,* but also that the rules and laws are at least somewhat *accurate*. Furthermore, if it were true that these generalizations are used pervasively, then it should not be problematic at all to name them.

Goldman and Gordon both furthermore criticize that the utilization and skilled application of this kind of rules cannot explain how children, who are – relying on empirical evidence like the false-belief task – thought to acquire the ability to attribute mental states to others correctly at the age of 5, make sense of others. Not only does this strategy seem way too sophisticated for children, TT in addition has no plausible story about how these laws should be acquired, as Goldman argues. It cannot be by cultural transmission, since folk psychological laws are, firstly, not spelled out explicitly very often, and, secondly, if they are, they are spelled out by philosophers. However, not many children have access to philosophers, as Goldman sarcastically comments. Neither can it be explained by personal construction, since it seems implausible that every child independently builds the same set of laws.

Along those lines, Gordon (1986) claims that the acquisition of the ability of mental state attribution comes with the acquisition of making assertions about the other's behavior in the context of practical simulation. This is, according to him, experimentally verified in the false-belief task. However, Gordon rejects the then-prevalent interpretation of the results that children who fail the test have not acquired a set of folk psychological laws. He argues that if the ability to attribute mental states is a matter of internalization of a set of rules at a certain point in development, this does not explain why children fail to attribute *false,* but not *true* beliefs. TT does not make any predictions about these semantic differences, but merely posits that children either possess or lack the required ability. Thus, if the child acquires ToM only by the age of 5, she should not be able to attribute *any* propositional attitude to others. Gordon further asserts that ST is better able to interpret the results of the false-belief task. In simulationist terms, the child simply has to overcome her egocentrical viewpoint as a developmental stage. Before this has not happened, she will rely on her own beliefs too heavily and thus not be able to attribute beliefs that are different from her own.

1.3.3. Putting Oneself in the Other's Shoes

As I showed in chapter 1.2.1., the debate about mental state attribution arose within the debate about the nature of mental representations. TT got off as a theory about the way mental concepts gain their meaning, but clearly developed to become a theory about mental state attribution to oneself and others. ST challenges TT concerning its assumptions about the *strategy* and *mechanisms* that underlie the attribution of propositional attitudes. Thus, it is fair to say that with the introduction of ST, the focus of the debate further shifted towards more specific questions concerning social abilities and away from general philosophical questions.

Additionally, in contrast to most versions of TT, ST does not depict mental state attribution as a metarepresentational skill, but rather as a *process-driven* capacity:

> They are 'putting themselves in the other's shoes' in one sense of that expression: that is, they project themselves into the other's *situation,* but without any attempt to project themselves into, as we say, the other's 'mind'. (Gordon, 1986, p. 162)

Gordon – one of the earliest pioneers of ST – describes simulation as projection, but importantly, he emphasizes that this process does not end up in a projection of oneself entirely. How does he arrive at a theory that on the one hand incorporates the claim that simulation draws on workings of one's own mind, but on the other hand states that we do not project ourselves into the other?

First, he addresses the need for a plausible account of how the mind works in general, which comes with the assumption that the workings of one's own mind form the basis for simulation. Gordon (ibid., p. 159) attempts to yield such an account when he begins his paper with a description of how we predict and understand our own behavior by "simulated practical reasoning". This practical simulation is the hypothetical decision what to *do* (hence, practical) without *actual* execution (hence, hypothetical). The same strategy of "hypothetico-practical reasoning" (ibid., p. 162) is also used to understand others. Yet an additional act is necessary to make the reasoning really about what *the other* will do or has done and not about what *I* will do or have done in a certain situation. Both a spatiotemporal shift in one's perspective and the adoption of the other's role are necessary for hypothetico-practial reasoning that arrives at an explanation of the other's behavior. In doing so, one draws on the "principle of least pretending" or "least effort principle" (ibid., p. 164), which means that the most realistic scenario is preferred, the one that deviates least from "what I myself take to be the world" (ibid.).

Importantly, in Gordon's version of ST, simulation does not necessarily involve a *comparison* to oneself, since this would restrict the scope to situations in which the person herself has been in before. Recall that Scheler accused proponents of the argument from analogy of drawing a picture of empathy that never actually arrives at understanding the other, but only oneself (see chapter 1.1.2.). Although not made explicit by Gordon, the introduction of the principle of least pretending avoids the criticism that inference on the basis of one's own mental contents does not yield any information about the *other,* but that its scope is by necessity restricted to yield information about *oneself.*

Going back to the quotation above, simulation appears as a mechanism that operates at the personal level. The metaphors and terms used to describe simulation tell as much; one puts oneself into the shoes of the other person, or one projects to the other what she herself would have done. However, Gordon (1986, p. 170) stresses that self-reported pretense may only be found in a small amount of cases and most of the crucial processes operate as 'offline simulation' at a sub-personal level:

> Since the system is being run offline, as it were, disengaged also from its natural output systems, its 'decision' isn't actually executed but rather ends up as an anticipation, perhaps just an unconscious *motor* anticipation, of the other's behavior.

Furthermore, it is hypothesized that the readiness and automaticity which characterize the simulation process may be due to a "prepackaged module" (ibid.) that is activated when we perceive others.

The conception of simulation that Gordon establishes here aims to offer an alternative account of the nature of folk psychology, namely, folk psychology *as* simulation. Although Gordon avoids to talk about representations or even metarepresentations, he concedes to TT that predictions of behavior are indeed often put in terms of propositional attitudes. Importantly, the author neither denies the functionalist assumption of proponents of TT that "common discourse about beliefs and other mental states presupposes that they enter into a multitude of causal and nomological relations" (ibid.), nor that these causalities can be couched in terms of lawful regularities that then serve as the basis for prediction of behavior. However, a too mechanical application of these generalizations is rejected. The generalizations that help to predict behavior are being used in *the context of practical simulation* rather than theorizing. In short, while TT states that the nature of folk psychology is a set of laws in the form of a theory, Gordon's version of ST claims that folk psychology draws upon a *process* of simulation, that may be taken offline and work at an unconscious level.

1.3.4. Constructing the Other – Simulation as Interpretation

While TT arose from the rejection of Cartesian claims of the centrality of the own mind, ST seems to go back to these assumptions (Gopnik & Wellman, 1992). Understanding and accessing one's own mind is fundamental for simulation to work, since it is the very basis for its crucial mechanisms. Thus, in rejecting the central claim of TT – that mental state attribution happens on the basis of a universally valid set of folk psychological laws that do not depend upon individual minds – the first-person perspective gains center stage again. As we shall see in the following, Goldman picks up this claim and constructs his version of ST around it.

In his view, the ascription of mental predicates involves the ascription of a specific *content,* hence the term *interpretation*:

> Thus, ascription of mental states, especially the attitudes, can be thought of as a matter of interpretation; and the strategy of studying the interpreter, in order to extract the conditions of mentality, or propositional attitudehood, may be called the *interpretation* strategy. (Goldman, 1989, p. 161)

Goldman aims to investigate this interpretation strategy, because he sees a proper description of what people actually do when they attribute or ascribe mental states and, even more importantly, how they arrive at the ascription of propositional attitudes, as central explananda of a theory of mental representations that has not been yielded by any other philosophical theory so far. His conception of ST, which shall be described in more detail now, targets to fill this gap.

Gordon's view that has been described in the previous section carefully works out how simulation may work at the sub-personal level. It also circumvents the criticism that by simulating, one might end up giving information about oneself rather than the other person. In contrast, Goldman (ibid., p. 169) seems to draw a cruder picture of the process:

> Rather, they ascribe mental states to others by pretending or imagining themselves to be in the other's shoes, constructing or generating the (further) state that they would then be in, and ascribing that state to the other. In short, we *simulate* the situation of others, and interpret them accordingly.

What is compelling in this quotation is that Goldman uses the word 'constructing' to describe the simulation process and thus clearly sides with a view that depicts it as a high-level, conscious process.

This also becomes obvious when we consider his claim that one advantage of simulation is that it is *introspectively* plausible. According to the author, we do find ourselves putting us into the mental shoes of another person and thereby predict or interpret her behavior. In other words, our phenomenal experience seems to fit with the claims of ST in the sense that we are indeed able to consciously think about what we would have done if we had been in the other's situation. Thus, in contrast to Gordon, the picture of simulation drawn by Goldman *prima facie* appears to describe the process as a high-level mechanism, operating at the level of and with the help of conscious thinking.

Further, Goldman postulates several constraints that must be fulfilled for simulation to work. Firstly, it is stated that the simulation process requires an *isomorphism* between the interpreter and the one whose behavior is being interpreted: "For a device to simulate a system is for the former to behave in a way that 'models', or maintains some relevant isomorphism to, the behavior of the latter" (Goldman, 1989, p. 173). This isomorphism must be present for both the *process* driving the simulation (which must hence be isomorphic to the process that drives the target system), and for the *initial states* of the simulator and the target system:

> Thus, if one person simulates a sequence of mental states of another, they will wind up in the same (or isomorphic) final states as long as (A) they began in the same (or isomorphic) initial states, and (B) both sequences were driven by the same cognitive process or routine. (ibid., p. 174).

The requirement of isomorphism is reminiscent of c_3 – similarity, a component of social cognition I postulated earlier. What they have in common is the idea that in order to project the states of oneself to the other person, they must share (some) features. However, Goldman's claim is much stronger in the sense that similarity would not be sufficient here; much rather, both the initial and the final states of two individuals must be the *same*. Such a strong claim faces the profound challenge of how self and other are to be distinguished, and how people may understand mental states of other people that are different from their own ones.

A further constraint concerns features of the content that is being attributed to the other person. In accordance with Quine and Lewis, Goldman claims that there must be constraints concerning the kind of contents that are attributed in order to account for the accuracy of prediction and interpretation and to explain why we tend to attribute one rather than another content. At this point, Goldman rejects the underlying functional notions of mental states, claiming that they cannot explain why some contents come more naturally than others and people hence end up to possess some kind of tacit knowledge. He argues that functions can be earned in non-biological, multiple ways and thus need not to be natural kinds. Thus, by assuming that the content of mental states is defined by their functional or causal role, no useful claim is being made about how this content is similar (and thus seems to come natural) in a variety of individuals.

Goldman now tries to circumvent these shortcomings and cover the relevant questions about content by making the assumption that since the *categories* and *processes* are grossly similar in the interpreter and the other person, this explains similarity of beliefs without making the set of beliefs too narrow:

> The simulation hypothesis assumes that the interpreter tends to impute to the interpretee the same fundamental categories as her own, or at least the same basic category-forming (and proposition-forming) operations. She also tends to project the same basic belief-forming processes. But these practices still leave room for wide divergence in belief content. (Goldman, 1989, p. 180)

Although the isomorphism constraint still holds, this conception leaves place for the interpretation and prediction of the behavior of non-conspecific beings, since it allows for a variety of belief sets with a variety of contents. As Goldman (ibid., p. 181) further notices, this version of ST exerts a kind of analogy in the manner of the argument from analogy: “It seems to impute to interpreters inferences of roughly the following form: ’If he is psychologically like me, he must be in mental state M; he is psychologically like me; therefore, he is in mental state M.’” Being well aware of the criticism that the argument has been exposed to, Goldman offers an answer to the question of how analogy can be true if it cannot explain how interpreters come to believe that the other is psychologically like herself. The solution lies in the rejection for the necessity to hold this belief explicitly. As Goldman argues, it is not necessary that the beliefs we operate on are held and represented consciously. He gives the example of perception, stating that it could well be possible that our perception relies upon Gestalt principles that are not represented explicitly by the perceiver. Yet they structure our perception.

A critical point of Goldman’s position is raised by Churchland (1989). He complains that ST cannot explain how we come to understand mental states that we have not experienced ourselves. According to him, Goldman’s version of ST only applies to situations in which the interpreter has herself been in before. This criticism makes sense, given the isomorphism constraint. If my initial state must be the same as yours in order for me to understand your situation, we have to have the same experiences. If you are in a state that I have never been in, there is no way for me to simulate your behavior, because the very basis is lacking.

In response, Goldman defends himself by claiming that theoretical inference is not denied entirely, and can still be used when simulation comes to its limits. Thus, whenever the required constraints are not fulfilled and simulation fails, we may still use theorizing to understand the other person. In doing so, he opens up the possibility of a hybrid account of mental state attribution or mindreading – a move that was of major importance in his later work, as we shall see in chapter 1.4.1.

1.3.4. Mirror Neurons and Low-level Simulation

While the views I have presented so far only speculated about the details of simulation, research from developmental psychology offered a new idea. Meltzoff and Moore (1977), for example, found that newborn babies show the tendency to imitate observed movements (such as tongue protrusion) short after birth. They thus proposed that the mechanisms underlying this ability are rather automatic and are based on a system than matches visual with proprioceptive information. In more detail, in order to imitate, an individual has to match the visual information about the observed movement with information about the proprioceptive or motor details about this movement. The idea of matching sits well with ST, since it also entails the usage of own knowledge (i.e., motor representations) for a social act (i.e., imitation).

In the early 1990s, a group of neuroscientists discovered a group of neurons in the brain of macaque monkeys that seemed to have exactly this function of matching representations of action execution with representations of action observation (Gallese et al., 1996). Mirror neurons, as this group of neurons has been called, were found to not only fire when the macaque executes an action, but also when he merely observes the same action, hence the conclusion that these neurons ‘match’ visual with motor information. The discovery of the so-called mirror neuron system had a great impact on the research field of social cognition and appeared to yield clear evidence for just the simulation process proponents of ST have been looking for. How exactly is that?

The claims of alleged functions of the mirror neuron system are put together in a very influential paper by neuroscientist Vittorio Gallese, who is part of the research team that discovered mirror neurons, and philosopher Alvin Goldman. In *Mirror neurons and the simulation theory of mind-reading* (Gallese & Goldman, 1998), they suggest that the discovery of the neural mechanism could yield profound support for ST. The authors suggest that the question how people attribute mental states to each other can be answered by adopting an evolutionary perspective and claim that this social ability is rooted in a simpler and phylogenetically older mechanism. In short, the mirror neuron mechanism is thought to form a primitive precursor of mental state attribution, here labelled as mindreading and defined as "[…] the activity of representing specific mental states of others, for example, their perceptions, goals, beliefs, expectations, and the like" (Gallese & Goldman 1998, p. 495). The mirror neuron system and its function to help action understanding thus is part of a representational capacity that enables mental state attribution.

As I have already shown, ST generally assumes that people use their own mental mechanisms as a basis for understanding others. The authors describe in more detail how this comes to pass, using the example of decision-making. When observing another person, the observer creates a *pretend* state that presumably caused or preceded the behavior. This pretend state then is fed into the observer's decision-making mechanism, which outputs the decision that she would have made. However, this decision is not acted upon, but 'taken offline' and instead used to predict the other's behavior. Mirror neurons could contribute a crucial bit to this process. Although it is stressed that their activity does not constitute a full-fledged process of simulation, they are thought to form a simulation heuristic that serves as a primitive precursor to the process. Externally provoked mirror neuron activity thus generates an action plan – just as does internally activated mirror neuron discharge. This action plan is not only taken offline or inhibited, but also 'tagged' as belonging to the person that is being observed. Their activity thus "seems to be nature's way of getting the observer into the same 'mental shoes' as the target – exactly what the conjectured simulation heuristic aims to do" (ibid., pp. 497–498).

Furthermore, the authors see the role of mirror neurons for simulation as evidence against TT. ST, in contrast to TT, predicts that the process of simulation involves mental mimicry:

> The core difference between TT and ST, in our view, is that TT depicts mind-reading as a thoroughly 'detached' theoretical activity, whereas ST depicts mind-reading as incorporating an attempt to replicate, mimic, or impersonate the mental life of the target agent (ibid., p. 497).

This claim yields, according to the authors, the possibility to empirically test whether ST or TT should be preferred. If it turned out that mind-reading indeed includes mental mimicry, this would certainly speak for ST, which predicts such a mechanism.

In summary, the simulation theory of mind-reading as spelled out by Gallese and Goldman claims that the finding of a class of neurons that code both action observation and action execution can be interpreted as strong evidence for ST. This is because mirror neuron activity, in functioning to match representations of observed actions with the own motor repertoire of the observer, generates a motor plan for the observed action that serves as a pretense state that can be attributed to the other.

The hype that soon centered around mirror neurons and their meaning for understanding others came with, but also contributed to, several changes in how social cognition was seen. Once viewed as a metarepresentational, high-level ability of (not too young) humans, mental state attribution became to be thought of as being enabled by low-level, sub-symbolic processes. In his later work, Gallese (2005) picks up the notion of simulation as depending on a low-level, bodily mechanism and elaborates on his theory of *embodied simulation*, which will be described in more detail in the next section. The growing tendency to understand simulation as the main mechanism for social cognition and as a mainly automatic, sub-personal process is interesting under several aspects.

Firstly, it comes with a general acknowledgement of the role of the body for cognitive processing both in the research field of cognitive and social cognitive science. This acknowledgment grew with the acceptance of 'embodied cognition' as a serious research field. A second point that is especially crucial for this thesis is that ST was introduced by psychologist Gordon and is now presented by a philosopher and neuroscientist. The version of ST I just presented is a prime example for interdisciplinary work and how it can illuminate each research field. In this case, neuroscience appeared to have found the biological basis for what psychologists and philosophers had previously described as simulation, yielding the conceptual framework for the empirical finding. I am convinced that such an interdisciplinary approach is indispensable when the aim is to find out more about such a complex phenomenon as social cognition and thus add another desideratum to the list:

d_2 – interdisciplinarity
Interdisciplinarity demands that a framework on social cognition considers theoretical and empirical findings from several disciplines and enables a dialogue between them.

Further, although the focus in the research field on social cognition shifted more and more towards low-level processes that involve bodily structures, the terminology at hand seemed to stagnate. The theory that has just been reviewed is a compelling example; the central claim of Gallese's and Goldman's paper is that the process of simulation crucially depends on a *low-level, sub-symbolic* system that usually processes *motor acts*. This clearly implies that understanding others heavily depends on understanding and representing their *motor behavior,* and thus that the shared basis upon which simulation then runs is as well *motoric*. However, the ability they refer to is dubbed mind-*reading,* suggesting that the process is exactly *not* one that implies bodily movements. The term much rather invokes associations to Fodor's (1975) language of thought hypothesis (LOTH), which depicts thinking and thought as a mental language. If this view is adopted and mental representations are thought to have a symbolic format and are syntactically organized, just as language, it indeed makes sense to call the process of understanding others mind-*reading*. However, it is questionable whether the usage of the term in Gallese and Goldman's work really aims to speak for LOTH. To the contrary it seems that they wish to contribute a view of social understanding that does not depend on explicit, symbolic representation, as we shall see in the next sub-chapter.

1.3.5. Embodied Simulation

The idea of embodied simulation (ES) is embedded in a broader, three-level description of understanding others; the *shared manifold hypothesis* (Gallese, 2001). It states that there are three vertically organized levels at which the processes that enable individuals to understand each other can be described. At the phenomenological level, there is *intentional attunement* (Gallese, 2004), that is, the phenomenal experience of sharing and being connected to another person. One level below Gallese describes ES as the *functional* mechanism underlying this experience, which in turn is constituted at the lowest level by the mirror neuron system in the primate's brain (Metzinger, 2009). Interestingly, Gallese (2005) describes the mechanism of embodied simulation to underlie not only action and emotion understanding of other people, but describes it as the functional means by which bodily awareness (i.e., of one's own and others' bodies) is encoded in space.

The explanatory scope of the shared manifold hypothesis and thus the mirror neuron system is – as described by Gallese – quite comprehensive. It is not only a theory about social cognition, but also about how a *neural* system can be described as the basis for the generation of meaning, thus being a candidate precursor for generating propositional or symbolic mental content. What we thus find here is an attempt to couch simulation into a multi-level description of understanding others, encompassing the implementational, functional, and phenomenal level of analysis. Since, as I will keep showing throughout this thesis, social cognition is such a manifold phenomenon, a multi-level analysis will prove useful for any theory of social cognition and thus shall be listed as one central desideratum:

d_3 – multi-level analysis
Multi-level analysis asks for a description of social cognition at several levels of analysis in order to capture its diverse components.

Let me now describe the idea of ES in more detail in order to get a more detailed grasp of Gallese's view.

The main claim of his paper *Embodied simulation: from neurons to phenomenal experience* (Gallese, 2005) is that both the representation of one's own body in relation to the external world and the representation of other bodies are enabled by the same functional mechanism, viz., embodied simulation. Movements are crucial for establishing spatial awareness of one's body since they allow for the encoding of positions in space relative to the individual. Action simulation plays an important role here. According to Gallese, seeing an object triggers the representation of potential ways to act upon it, thereby establishing a *simulated motor plan* that hinges on the predictive computation of limb position, consequences of movements, etc. This process of simulation depends on sensorimotor integration since proprioceptive information needs to be merged with visual information (of both the environment and one's visible body parts)[13] in order to generate a simulated potential action (Gallese, 2005). Thus, when an action is executed, its consequences are computed as well. The same happens when observing the actions of other individuals – because both processes are *simulation* processes. The properties of the underlying neural network reveal further features of this functional mechanism. Drawing on data from research on macaque monkeys, Gallese (ibid., p. 33) claims that mirror neurons show properties that are similar to "symbolic properties so characteristic of human thought", since the content of their multi-modally driven activation is similar to the conceptual content of goal representation. The evidence that Gallese brings in are studies that show that mirror neurons in the macaque brain have audio-visual properties.

[13] Gallese (cf. 2005, pp. 27-29) illustrates how this process can be disrupted in patients with lesions to sensorimotor circuits.

Kohler and colleagues (2002) showed that mirror neurons fire both when the monkey is presented with an edible object (in this case, a peanut), but also when she merely hears a sound associated to the object (e.g., breaking the peanut).[14] The activation pattern of mirror neurons is described as possibly being equivalent to the use of verbs, because the pattern remains the same, no matter the operational context the 'verb' is presented in (*hearing* someone breaking a peanut or *seeing* someone breaking a peanut) (Gallese, 2005, pp. 33–34).[15] The upshot of this is the following: The activation of mirror neurons – due to their specific properties – generates representational content about the *meaning* of actions via simulation. Since this mechanism is recruited for both one's own and others' actions, it forms a common basis for understanding oneself and others.

According to Gallese, the applicability of ES does not stop with action understanding. Additionally, the mechanism realized by the mirror neuron system is also deemed to underlie the understanding and recognition of emotions:

> When I see a given facial expression, and this perception leads me to understand that expression as characterized by a particular affective state, I do not accomplish this type of understanding through an argument by analogy. The other's emotion is constituted and understood by means of an embodied simulation producing a shared body state. It is the body state shared by the observer and the observed that enables direct understanding. (ibid., p. 39)

The quotation shows that Gallese also relates emotion recognition to the conception of bodily awareness as construed within a process of sensorimotor integration and thus aims to get rid of the need for analogical inference. What is most interesting here, it seems to me, is the claim that not only emotion recognition, but also the emotion *itself* is claimed to be constituted by ES. Since Gallese views the mirror neuron system as the basis of ES, this means that emotions and emotion understanding, actions and action understanding, concept formation and high-level thought are enabled in virtue of neurons with mirror properties.

What is more, Gallese (ibid., pp. 39-40) attempts to also incorporate a phenomenological view:

> Were we adopting this [phenomenological] perspective to frame social cognition, we could say that the self-modeling functional architecture of the alive body scaffolds the modeling of

[14] Another study that is named as supporting the claim is Umiltà and colleague's (2001) work that showed mirror neuron activation in monkeys even when the crucial part of the observed action was occluded. This finding is taken as evidence that mirror neurons encode *whole* actions, including their goals, and not mere movement sequences.

[15] For a more detailed proposal of how mirror neurons could underlie concept formation, see Gallese & Lakoff, 2005.

> the intentional relations of other individuals. The multimodal dynamic model of our body as of a goal-seeking organism, brings about the basic representation architecture for the mapping of intentional relations.

If it has not been obvious before, it should now be clear just how comprehensive the theory of ES aims to be. Spanning from the neural to the phenomenological level of description it offers explanations on how the representational or functional level in between is not only determined by a specific property of neurons, but also by our very bodies. This is an important step for ST. It appears that with the discovery of mirror neurons, the research community found the basis for simulation and thus understanding other people:

> Our seemingly effortless capacity to conceive of the acting bodies inhabiting our social world as *goal-oriented persons* like us depends on the constitution of a shared meaningful interpersonal space. This shared manifold space can be characterized at the functional level as embodied simulation, a specific mechanism, likely constituting a basic functional feature by means of which our brain/body system models its interactions with the world. Embodied simulation constitutes a crucial functional mechanism in social cognition, and it can be neurobiologically characterized. (Gallese, 2005, p. 42)

In characterizing our ability to understand others without effort, ES is depicted as a process that occurs automatically[16] and without the need to consciously initiating it. It is thus thought to enable us to understand others effortlessly and without the need to form theories or other conceptual reasoning in the first place. This sub-personal mechanism, which is neither controllable nor accessible, ultimately results in the personal level experience of being 'intentionally attuned' to others. Such a view has further important implications.

First, as Gallese stresses, emphasizing the automatic nature of ES does not mean that we are not capable of or never use more sophisticated strategies. While he asserts that humans are able to take a detached perspective and reason about other people's behavior, he emphasizes that the most *fundamental* capacity to understand others remains the low-level mechanism he dubs ES.

Secondly, note that Gallese calls our social abilities *seemingly* effortless and at the same time describes a rather complex underlying mechanism. This is of major importance for any multi-level analysis. In making this statement, Gallese dissociates the phenomenological from all other levels of description. For just because understanding others *seems* to come without an effort, does not imply that there is not a lot of 'effort' at *other levels* in the sense of complex architectures that bring forth this experience.

[16] In a footnote, Gallese (2005, p.44) qualifies that 'automatic' in the case of ES means "obligatory".

What we find here, to briefly sum up, is a comprehensive, multi-layered perspective. It seems that we found the basis for simulation and are able to relate the mechanism to its neural components. At the same time, however, we still find ourselves forming theories about the other person. The question therefore still stands: ST or TT?

1.4. The Mindreading Debate

The debate that focused on the question whether it is simulation or theorizing that is (mainly) used to understand other minds will here be called the 'mindreading debate'. Both ST and TT claimed to have found the basis of 'mindreading', where the term refers to the ability to attribute mental states to other people and thereby understand, predict and interpret their behavior.

There are several interesting developments in the mindreading debate that I will discuss in this chapter. Most compelling, it seems to me, is the shift towards hybrid theories of ST and TT. While the primary question for a long time seemed to be if it is simulation *or* theoretical inference that makes us social, researchers quickly picked up the idea that it might be both. The discussion then rather focused on asking which of these mechanisms are developmentally prior and quantitatively prevalent, that is, which arise earlier in development and are used most often in adult life.[17]

Taking this development as a starting point, I will take the following steps:

(1) To begin with, I claim that forming hybrid versions of mindreading theories has been possible because ST and TT share a common set of metaphysical background assumptions. These will be described in chapter 1.4.1.

(2) In chapter 1.4.2., I show another development that is related to forming hybrids, namely the move to 'go sub-personal'. In more detail, both ST and TT start to describe the mechanisms they foster at the sub-personal level. Consequences of such an endeavor and its usefulness will be discussed.

[17] Note how this sub-debate is different from the question of which mechanisms are at play in social cognition *in general*, no matter how rarely they are recruited.

1.4.1. The Common Background of ST and TT

Before depicting the *integration* of ST and TT into hybrid accounts, I will first discuss the common background of both theoretical accounts which made this theoretical shift possible. ST and TT as philosophical theories of social cognition arose at a time when the mind was thought to be a computational, representational device that was to be found inside of the skull of an individual. That is, the mind is assumed to be a functional structure that is locally realized in the brains of individual organisms. Bodily and environmental structures were attributed an at most enabling or causal role for internal mechanisms. This internalist picture of mental processes served as the theoretical background into which both theories were couched. So far, I have depicted the historical growth of both theories, which is certainly part of the reason ST and TT have been spelled out in a cognitivist set of assumptions. However, there are deeper systematic reasons why both theories (at least implicitly) are based upon this specific set of metaphysical assumptions about the mind.

Cognitivism states that the task of the brain is to grasp what is going on in the outside world by *internally representing* it. Other individuals are part of this world outside one's own mind and it thus follows that the causes for their behavior need to be inferred by internal representational processing, too. Since the brain is taken to be the only mental organ, (social) cognitive processing is located in the individual's head. ST and TT nicely fit into that picture, since both simulation and theorizing are inference processes that function to attribute mental states to other people by inference and are implemented by specific neural mechanisms. In other words, both theories assume that in order to make sense of a perceived behavior, to predict future behavior and to act accordingly, an *inferential* mechanism needs to be activated in between perception and action (e.g., Gallagher & Zahavi, 2008). This furthermore implies the assumptions that (1) behavior is caused by mental states, and (2) since these mental states are not directly accessible through perception they (3) need to be inferred.

The role of perception in mindreading accounts is peripheral in the sense that the main work load is done by cognition. Perceiving the behavior of other people merely yields the input which is then processed by cognition. The resulting action towards other people then is the behavioral output. Hurley (1998, p. 401) describes this view of the mind as a "sandwich" or an "input-output device". This reflects the assumption that perception yields sensory inputs that then are processed by cognition as the central part, producing a particular output, action.

These assumptions clearly lead to a focus on cognition as the crucial research target, while perception and action are seen as rather peripheral.

The picture that emerges here depicts social cognition as an inferential, internal and representational mechanism. What is more, it is a rather individualistic view, for the context in which most social cognitive processing takes place – namely, interaction – does not play a noteworthy role. This is also reflected in the empirical paradigms that have been used to investigate mindreading. Most experimental set ups are rather static without any interaction between individuals. However, it is important to note that this is also due to methodological restrictions. Almost every method that can be used to image brain mechanisms underlying a cognitive phenomenon require the participant to keep still and move as little as possible. At the same time, it appears that this restriction did not appear as problematic to most researchers, since mindreading was seen as something that takes place inside an individual's head.

This common background made it possible to assume that mindreading may involve both simulation and theorizing. While it once seemed to be an either-or-question whether mental state attribution is achieved by running a simulation of one's own mind *or* by drawing on a set of theoretical entities, hybrid versions arose, stating that mindreading involves both mechanisms, depending on which task is to be achieved (Apperly, 2008; Nichols & Stich, 2003). Stich and Nichols (1992), for example, suggest that off-line simulation could well draw on a tacit theory. Goldman (2006, p. 43) endorses a hybrid account of mindreading, depicting "a number of ways to blend simulation and theorizing elements into a mosaic of mindreading possibilities." In these hybrid accounts, it is often assumed that theorizing is a rather sophisticated and high-level process that is only available for a specific subgroup of individuals. Simulation, in contrast, is seen as the 'baseline' of mindreading, encompassing low-level, unconscious processing.

However, this is not always the case. Throughout the debate, there have been attempts to depict theorizing as a sub-personal process, or to describe simulation as something that takes place both at the sub-personal and personal level. These diverse versions were made possible by the move that I will describe in the next subchapter, namely, to 'go sub-personal' (Slors, 2012).

1.4.2. Going Sub-Personal

As for ST, early proponents of this account such as Lipps or Gordon have already considered the possibility that simulation is a rather low-level and subconscious process. This idea has found empirical support with the discovery of mirror neurons. One criticism of TT has been that the ability they advocate is too sophisticated and thus excludes non- and pre-verbal individuals. To counter this accusation, there have been attempts to postulate that the theory that is being drawn on in making sense of others is to be seen as 'implicit' (e.g., Bermúdez, 2003).

At this point it is important to acknowledge a difference in making sense of the claim that a particular ability or mechanism 'is sub-personal'. Consider the two following strategies to do so: On the one hand, we may try to explain how a high-level cognitive skill builds on or is constituted by lower level mechanisms, decomposing the ability into its 'dumber' parts and to find out which underlying skills there are that enable this more sophisticated one. This is a quite common endeavor in the cognitive sciences.

On the other hand, one can claim that a personal-level mechanisms is *to be found at* lower levels. One strategy in (social) neuroscience is to find correlations between the activation of a brain region, functional or cognitive mechanisms and behavioral phenomena. This is what Frith and colleagues, to name one example, do when they claim that they have found a brain region that most likely correlates with ToM or mentalizing abilities (Frith, 1999). Lesion studies can even reveal causal relations. This strategy takes a phenomenon – say, mental state attribution on the basis of a set of folk psychological laws – and tries to find its lower-level counterparts,[18] but without 'degrading' the phenomenon itself to a sub-personal process. This strategy of finding components that play an enabling or constitutive role for the occurrence of a higher-level phenomenon is different from qualifying the phenomenon *itself* as one of these components. That is, however, what 'going sub-personal' in the mindreading debate often involves.

[18] Consider, for example, 'homoncular functionalism', which Dennett (1994, p.240) describes in a self-portrait: "The best known instance of this theme in my work is the idea that the way to explain the miraculous-seeming powers of an intelligent intentional system is to decompose it into hierarchically structured teams of ever more stupid intentional systems, ultimately discharging all intelligence-debts in a fabric of stupid mechanisms […]. Lycan (1981) has called this view homuncular functionalism."

Both simulation and theorizing are no longer described as personal-level abilities that individuals consciously apply for understanding others. Gallese's (2005, p. 33) theory of embodied simulation, for example, states that simulation is a functional, sub-personal processes which is "embodied by" mirror neurons. Nichols and Stich (2003), defending a hybrid version of ST and TT, differentiate between simulation and theorizing by describing the latter as an 'information-rich' process. This process is further characterized as drawing on "a rich set of mental representations containing substantial information (or, sometimes, *mis*information) about mental states and their interactions with environmental stimuli, with behavior, and with each other" (ibid., p. 102). As Herschbach (2008) correctly concludes, this description allows for many sub-personal processes to count as theoretical, also because the authors do not determine the format of these mental representations as 'theory-like', i.e., symbolic or conceptual.

Along the lines of the empiricist account of TT as described by Gopnik and Wellman (1992), Gopnik and Meltzoff (1997) describe structural, functional and dynamic features of theories which are thought to also be features of the child's theory of mind. A closer look at their definition of a theory reveals that they are by no means committed to a personal-level account of TT: "A person's theory is a system that assigns representations to inputs just as one's perceptual system assigns representations to visual input or one's syntactic system assigns representations to phonological input" (ibid., p. 43).

Blackburn (1992) has criticized the description of theories in sub-personal terms as promiscuous, leaving the terms they use inexpressive. While it does not seem implausible to characterize both simulation and theorizing as sub-personal processes if couched in the right terms, it does add confusion to the conceptual landscape. Since this shift to the sub-personal level has not at the same time erased the personal level usage of the terms, both simulation and theorizing are left with an ambiguous meaning.

Furthermore, there are several ways to characterize these versions of the terms. Take Goldman's hybrid theory of mindreading as a famous example. In his newer work, he describes mindreading to involve both high-level and low-level simulation. While the latter is clearly characterized at the sub-personal level as automatic, fast and mainly unconscious, high-level simulative mindreading has

> […] one or more of the following features: (a) it targets mental states of a relatively complex nature, such as propositional attitudes; (b) some components of the mindreading process are

> subject to voluntary control; and (c) the process has some degree of accessibility to consciousness. (Goldman, 2006, p. 147)

In this statement, however, it remains unclear whether and how the distinction between low-level and high-level simulative mindreading maps onto the one between sub-personal and personal level processes. This point is described by Hurley (2008, p. 762) in more detail:

> Indeed, my initial response to Goldman's high ⁄ low distinction was to regard it as a version of the personal ⁄ subpersonal distinction. But it doesn't take long to see that this reading isn't right; Goldman makes no explicit appeal to the personal ⁄ subpersonal distinction in characterizing high vs. low, nor does he invoke standard criteria for personal vs. subpersonal levels. Personal vs. subpersonal seems to cut across Goldman's distinction: for example, experiencing and understanding emotions are low level processes but can certainly be described at the personal level. For these reasons, I've found Goldman's distinction elusive and would welcome clarification."

Hurley picks up an important point that is symptomatic for some difficulties of the mindreading debate, namely that 'going sub-personal' and the development of hybrid theories has added diversification and thus equivocation to the theoretical and terminological basis the debate.

At the same time, those moves involve an acknowledgement of the importance of low-level, non-conscious processes for social cognition. This has surely been an important step for philosophical theories that aspire to be empirically informed and plausible. It appears rather odd, however, that the shift to lower levels seems to find only little reflection at the terminological level. Instead of finding or using terms that allow for a vertical explanation of a given phenomenon in a non-equivocal way, participants of the debate have chosen to simply re-use many of the central terms. The tendency to stick with a rather conservative terminology furthermore shows itself in the way the inferential mechanisms between perception and behavior are described, as will be described in the next section.

1.5. Evaluating the Debate

In this subchapter, I deal with criticisms that mindreading approaches have faced during the past decades. Since these led to important and influential discussions that then caused what some like to call the 'interactive turn' (Gallotti & Frith, 2013; Overgaard & Michael, 2013), I find it worthwhile to examine whether this is a fruitful criticism.

(1) In chapter 1.5.1., I evaluate the terminological landscape of the mindreading debate and ask whether the terms and concepts used are appropriate. As it will turn out, most

terms reflect rather outdated philosophical intuitions instead of expressing the current state of the debate.

(2) Section 1.5.2. asks where the mindreading debate is leading and questions its efficiency. Reasons for this criticism will be presented and evaluated.

(3) Lastly, it will be discussed whether mindreading should be the central explanandum in the research field of social cognition. I present several critical points which cast doubt on the view that mindreading as depicted by ST and TT is our most fundamental social ability.

1.5.1. The Terminological Landscape

In the following, I show how central terms that are mostly being used – ToM, mindreading and mentalizing – are all reminiscent of the long discarded philosophical conception of TT and thus do not reflect the trend to go sub-personal.

Let me begin this section by pointing to one source of confusion right away, namely the relation between TT and the term 'ToM'. In doing so, I will first quote definitions that capture the consensus in the literature on how to conceive of the theory and the term:

> Theory–Theory (TT) accounts propose that theory of mind abilities are constituted by a set of concepts (belief, desire, etc.) and governing principles about how these concepts interact (e.g., people act to satisfy their desires according to their beliefs). The proposed status of these concepts and principles varies widely, from symbols and processing rules in sub-personal Language of Thought (for a discussion see Stich & Nichols, 1992), to a set of personal-level notions and hypotheses to which we have explicit access […]. (Apperly, 2008, p. 268)

> Theory of Mind (ToM) is the cognitive achievement that enables us to report our propositional attitudes, to attribute such attitudes to others, and to use such postulated or observed mental states in the prediction and explanation of behavior. (Garfield, Peterson & Perry, 2001, p. 494)

The conceptual proximity and the historical development of both the theory and the term give the impression that ToM is inextricably linked with TT. Thus, when philosophers and scientists talk about an individual having a ToM, it seems natural that they are proponents of TT. However, this is not the case. Proponents of TT claim that the source of knowledge that serves as a basis for mental state attribution is indeed a *theory* (which can be implicit or explicit, or both) – thus the label 'theory-theory'. ToM, in contrast, is a term that is used very loosely and has been interpreted in countless ways.

It is also ambiguously used to describe both a *mechanism* that underlies mindreading as well as an *equivalent* to mindreading. In the latter case, ToM refers to the ability that itself recruits several mechanisms. Interestingly, when described as a cognitive achievement, ToM is also used in theoretical accounts that actually deny TT (or its prevalence):

> Although historically it has been seen as distinct from simulation, theory-of-mind ability, broadly construed, encompasses several distinct strategies and several neural regions with a single goal: to understand the internal states that predict the behavior of other people. (Adolphs, 2009, p. 706)

Such a wide conception of the term does not commit those who use it to be proponents of TT. At least they do not have to subscribe to seeing social understanding as being accomplished *solely* by theoretical inference. One thus has to be careful not to equate the term with the theory – since the latter indeed depicts the basis of mental state attribution as theoretical knowledge.

The question then is: If scientists as well as philosophers agree that having ToM does not *necessarily* involve holding a *theory,* either at the personal level or sub-personal level, then why does the research community stick with the term so persistently? There are several possible answers to this question.

Slors (2012) argues that the strong intuition that ToM is ubiquitous stems from mistaking the *model* of ToM with the *actual* social cognitive mechanisms. Similarly to what Dennett (1971) describes as taking the intentional stance[19] – the *strategy* which humans apply to ascribe beliefs and desires to intentional systems – ToM as a model provides a way to *talk* about reasons of behavior without reflecting actual causal routes that lead to a certain behavior. In the words of the author:

> My proposal, however, is to take the intentional stance idea one level up: the cognitive mechanisms involved in understanding others and navigating the social world can best be tracked and understood for practical purposes in terms of the attribution of beliefs and desires, i.e. the application of a ToM. But that is not to say that these cognitive mechanisms use or implement a ToM. ToM is a model of ubiquitous social cognitive mechanisms that would otherwise be intractable. Since the model provides our ways of thinking and talking about such mechanisms, it may easily be mistaken for the real thing. This explains the intuition that ToM is ubiquitous. (Slors, 2012, p. 11)

[19] "The intentional stance is the strategy of interpreting the behavior of an entity (person, animal, artifact, or whatever) by treating it *as if* it were a rational agent who governed its "choice" of "action" by a "consideration" of its "beliefs" and "desires." (Dennett, 2013, p. 59)

What ToM-talk expresses, one could say, is a confabulation about what people think are the reasons for their own and others' behavior.

Another reason for the inclination to stick with the term may be that is has become an idiom (i.e., everyone 'just knows what is meant by ToM') and that is does not necessarily involve a theory of any kind. While an idiomized usage of a term is not a problem *per se,* things get fuzzy when the term is still being used in its literal sense as well. It is not obvious whether someone expresses her conviction that ToM indeed is related to theories in one sense or the other (e.g., Saxe, 2005) or whether it is simply an all-catch term that as well could imply simulation or any other mechanism. It thus seems that ToM is not 'idiom enough' to serve as a helpful term, since it still is being used in its original version.

The term 'mindreading' is used to describe the ability to understand what is going on in another person's mind and most often is referred to as a synonym for mentalizing or ToM. The term is problematic in two ways. First, it sets the focus on the mind, thus systematically ignoring a possible role of the body. It therefore clearly enforces a cognitivist notion that neglects the embodiment-debate entirely. Secondly, although the concept does not suggest a link to either ST or TT as strong as ToM, in its literal sense it still shows a tendency towards the latter. It does so in suggesting that a mind is something that can be read which presupposes it to be in a specific, viz., *readable* format. There is one philosophical theory of the mind that explicitly puts forth such a view, as I have already pointed out. Fodor's (1975) language of thought hypothesis (LOTH) starts with the premise that thinking is based on a mental language ('mentalese'). Thoughts and thinking, according to the philosopher, resembles language and talking in an important way: they both have a syntactic structure. Mental representations thus have a symbolic, syntactic format. Conceived of mental states like this, the notion of *reading a mind* seems quite plausible. Furthermore, LOTH clearly relates mindreading to linguistic abilities which in turn play the central role in 'original' versions of TT as described by Sellars and Lewis. However, since LOTH and the view of the mind as a syntactical and linguistic machine does not find many representatives nowadays anymore, it can be assumed that 'mindreading' is almost always used in its metaphorical sense. Again, the metaphoric usage of concepts does not have to be a problem. What it can be symptomatic of, though, is the lack of better, more accurate terms. This seems rather likely in the case of social cognition.

In chapter 1.2.6. I showed how the term 'mentalizing' has been introduced by Frith and colleagues as a synonym for ToM. While it has originally been described as a metarepresentational skill, there seems to be a silent agreement that this characteristic can be dropped. In current works, it remains unclear whether 'mentalizing' is referred to as a meta-representational ability or not.

In one of their papers, Frith and Frith (2012, p. 289) define mentalizing in the glossary as the "[…] implicit or explicit attribution of mental states to others and self (desires, beliefs) in order to explain and predict what they will do." As such it can refer to either a TT-bound view or the general inferential capacity in between perception and behavior that as well may recruit simulation mechanisms.

Hohwy and Palmer (2014, p. 172) use the word in order to describe general social cognition and additionally introduce the term of "meta-mentalising". That term is used to describe an ability which is applied when social situations get too complex to be accounted for by 'normal' mentalizing.

The many varying definitions of the term make it hard to parse which of them is valid in a particular context. Both the terms 'mindreading' and 'mentalizing' are often preferably used instead of ToM "because the former avoids begging the question of whether the capacity is to be explained in terms of the possession of a theory" (Overgaard & Michael, 2013, p. 2). However, a diachronic view of the concepts suggests that understanding others is an ability that draws on theorizing or at least depicts workings of the mind in a somehow linguistic format. That ST and TT indeed exhibit a rather 'old-fashioned' picture of social processes does not come as a surprise, given their historical development.

Let me briefly summarize the findings in this chapter. I showed that not only has the terminology of the mindreading debate lost its expressiveness by being used in different contexts so that a disambiguation of the term's meaning is only possible if made explicit by the user. I also showed that since central concepts still are strongly correlated to old-fashioned philosophical theories and intuitions, they do not reflect the trend to investigate social cognitive processes in sub-personal terms. The terminology that is available for researchers of social cognition thus seems not only outdated, but also inappropriate to enable a meaningful dialogue. What this leaves us with is the need for a terminology that is differentiating, i.e., able to properly refer to a skill, mechanism or ability without the need of extra disambiguation. I thus formulate the following desideratum for a theory of social cognition:

d_4 – terminological consistency
Terminological consistency postulates the need for expressive terms that are used in an unequivocal manner.

This will increase the descriptive power of a terminology and facilitate communication between researchers. For it is easier to understand the other's theoretical and empirical findings, when they are described within a common language.

1.5.2. A Debate Leading Nowhere?!

Let me now describe why one could say that the mindreading debate has failed to yield a sound and useful theoretical framework for social cognition. Apperly (2008) takes a critical look at both ST and TT and claims that although they set the agenda for much debate in the interdisciplinary research field and have shaped the conceptual and theoretical landscape profoundly, they have failed to offer suggestions on how to find out whether a specific social process involves simulation or theorizing. Despite the fact that there is consensus that mental state attribution recruits both simulation and theorizing, the author states that it is still important to find out which of the two actually underlies a particular social mechanism. However, as he exemplifies in his work, there has been little success in this endeavor.

One reason he identifies is that for almost every empirical finding, there is a version of *both* ST and TT that does the job in explaining it: "In many cases, behavioral evidence has proved inconclusive at discriminating ST from TT, because although data may allow some versions of ST or TT to be excluded, other versions of either theory are able to explain the findings." (ibid., p. 270) Another reason is that different studies operate on different concepts. Apperly presents several paradigms which attempt to identify simulation mechanisms by comparing self- to other-directed processes. Since it is claimed that simulation draws on models of one's own mind – while theorizing involves sets of abstract concepts and rules – simulation processes should recruit neural mechanisms that are used for both self- and other-directed processing (cf. Apperly, 2008). Because these studies all use different concepts of 'self', they cannot be said to investigate the same thing and are thus not comparable to each other. The author concludes that ST and TT have not been a too helpful framework to interpret empirical data and that "[o]n the basis of the current literature it seems possible that these theories will in fact become redundant as new findings about ToM motivate the development of new models based upon well-characterized cognitive and neural processes" (Apperly, 2008, p. 281).

Although ST and TT are influential theoretical accounts and although the debate has been going on for quite a while, there are little conclusive results or useful conceptual distinctions that would help to specify the mechanisms that underlie a specific social phenomenon. It is thus questionable where this debate is leading and whether it is asking the right question. In the next section, I will pick up this critical point and show how the debate may have focused on a too narrow explanandum.

1.5.3. Mindreading – The Central Explanandum for Social Cognition Research?

As Overgaard and Michael (2013) extract, there have been two ways of criticizing mindreading-approaches. The first is to reject *accounts of mindreading,* i.e., to criticize the way ST and TT describe and analyze the ability. This does not involve rejecting mindreading as the *central explanandum*, which is the second sort of criticism ST and TT have faced. While I mainly deal with the first in later chapters of this thesis, I will now focus on the latter.

The claim that mindreading is the central *explanandum* entails the assumption that mindreading is the central *ability* that is at the core of social cognition, which is what has been criticized extensively. However, according to Overgaard and Michael, not all participants in the mindreading debate describe the attribution of mental states as the only game in town and indeed are aware of the fact that quite many social situations can be dealt with without mindreading. At the same time, so they argue, it is false to reject the view that we *never* use inferential routines to make sense of others and also state that not all mindreading processes can be reduced to more basic, non-inferential forms of social cognition.

Although most critics would agree that – to whatever minimal extent – people sometimes need to recruit mindreading to understand others, their main point is that ST and TT depict the central social ability as observational and detached and thus pervert the ‘real’ picture of social cognition. It is claimed that our social life is much more complex and dynamic than described by mindreading accounts, which are accused of describing social cognition as *spectatorial, detached, individualistic, and disembodied*. Hence they are deemed unable to explain the ‘real deal’ of being social: interaction (e.g., Fuchs & De Jaegher, 2009).

While Overgaard and Michael (2013) show that ST and TT are not committed to strong detached and spectatorial views, but that this impression is rather due to the methodological boundaries researchers have to work with, things are more complicated when it comes to individualism and embodiment. These are issues at the very core of decisive questions of the debate, since they require choices of what kind of theory social cognition needs and what is aimed to explain.

Concerning individualism, one has to be careful to get the criticism straight and not to allege too strong accusations. The notion of an individualistic explanation of social cognition refers to the view that all relevant processes can be found in one individual and do not need to take external factors into consideration. First, note how this is a weaker claim than those that can be found in internalist explanations. It is one thing to claim that all relevant processes of social cognition are located within the skull of an individual. However, the view that all relevant processes can be attributed to the individual does not *necessarily exclude* the possibility that external processes *play a role for the individual*.

It could thus be possible to think of an externalist, yet individualistic account of social cognition. What critics of ST and TT claim is that individualism cannot be *the only* way to explain social cognition. They do not state that individualistic explanations are to reject *entirely*,[20] since internal and individual processes are considered crucial – but only to a certain degree (De Jaegher & Di Paolo, 2013).

When it comes to embodiment, the degree to which mindreading abilities are described as embodied has been criticized to be too small. Fuchs and De Jaegher (2009, p. 468), for example, accuse mindreading approaches to describe a "sender-receiver relation between two Cartesian minds" and that "[e]ven though simulation theories increasingly include the body in the modelling of others, they still do not take into account the reciprocity of embodied agents." Admittedly, the mindreading views that include the body do so in a rather weak sense that is fully compatible with old-fashioned functionalist and cognitivist views of the mind. For example, 'embodied' simulation (Gallese, 2005) is called that way because the crucial mechanism is thought to draw on a *model* of the body. This model, however, is represented in the brain of the observer, which leaves the crucial processes inside the skull and *not* inside the non-neural body.

Goldman and De Vignemont (2009) ask the question *Is social cognition embodied* in their same-titled paper and answer that the most useful way to view social cognition as an embodied process is to state that social cognitive processes exploit *representations that have*

[20] "We never suggested that individual cognitive performances are not relevant to some forms of social cognition." (De Jaegher & Di Paolo, 2013, p. 1)

a bodily (i.e., motoric, somatosensory, affective or interoceptive) format.[21] These bodily representations play a causal role for some social cognitive processes, as they argue showing evidence from lesion studies.[22] One famous example of non-pathological social cognitive processes that can be said to be embodied because of the format of representations they involve are motor representations that then form simulation heuristics for action understanding. Importantly, the authors state that not all social cognitive processes are embodied and that it is still open to discussion whether particular mechanisms causally depend on bodily representations or not. While low-level mindreading (as described in Goldman's hybrid account, see above and Goldman, 2006) thus can be said to involve representations in a bodily format, high-level simulation processes are more plausibly described as using representations in conceptual formats. The differentiation of kinds of representations thus also yields a way to distinguish a range of social cognitive processes. This description of social cognition endorses a deliberately weak and 'local' notion of embodiment and yet claims to stand in contrast to classic cognitivism. The latter would not, so the authors argue, assume the rather low-level nature of representations. Goldman and De Vignemont (2009) promote their "tame" position as fruitful for both science and philosophy. This exemplifies the reluctance one finds in the research field to accept radical positions of embodiment, some of which will be presented in the next chapter.

Taken together, there are several valid points of criticism which leave mindreading accounts unfit for providing a sound theoretical basis for social cognition. Most importantly, neither ST nor TT include components of the phenomenon such as embodiment and interaction sufficiently. It still appears that these aspects are thought to function in the periphery, instead of fulfilling a crucial role for social cognitive processing. Both of these components, however, are indeed important for establishing social relationships, as will be shown in the next chapter.

[21] Although the *content* of representations plays some role for defining them as bodily, it claimed that this is not enough, since one could think of representations that have one's own body as a content but are purely conceptual and amodal in format (Goldman & De Vignemont, 2009).

[22] "Patients with selective impairment in emotion experience have a matching selective impairment in recognizing emotional facial expressions in other people, whereas they have a preserved declarative knowledge about the relevant emotion […]. So, normal subjects, who have no such lesions, must be using their own emotion experience – involving a B-format – in recognizing the emotion of someone they observe." (ibid., pp. 156–157)

2. Phenomenology and Enactivism: The 'Phenactivist' Approach to Being Social

> *For we certainly believe ourselves to be directly acquainted with another person's joy in his laughter, with his sorrow and pain in his tears, with his shame in his blushing, with his entreaty in his outstretched hands, with his love in his look of affection, with his rage in the gnashing of his teeth, with his threats in the clenching of his fist, and with the tenor of his thoughts in the sound of his words. (Scheler, 1912/1973, p. 254)*

In his famous quotation, Scheler criticizes the argument from analogy which was presented in the previous chapter. His point is that there is no perceived effort to understand other people, that forming analogies is not part of our phenomenological repertoire. Instead, as becomes obvious in his words, we appear to get an immediate grasp of the other person. As a phenomenologist, Scheler zeros in the experiential quality of social encounters. His view can be seen as the locus classicus of what has later been dubbed 'direct perception' and which builds the bedrock of modern phenomenological theories of social cognition. Their criticism is also reminiscent of Scheler's; mindreading is seen as an almost fully redundant process that can be replaced by direct perception as a mechanism for social cognition.

The general goal of this chapter is to describe both the traditional and modern phenomenological perspective, and to show how this already profoundly different picture of social understanding is extended to enactive theories.[1] Enactivism puts much emphasis not only on the body, but especially on interaction as a potentially constitutive element of social cognition. The difference between enactive and phenomenological theories seems to boil down to the explanatory scope. While enactivism explicitly claims to offer a radically different alternative to cognitivism and thus to build a proper account of cognition (Varela, Rosch & Thompson, 1993), phenomenology is mostly seen as a description of experiential phenomena (Gallagher & Zahavi, 2008).

My main claim in this chapter is twofold. First, I see considerable shortcomings in both theoretical strands concerning their empirical plausibility and interdisciplinary validity. Secondly, though, I aim to show that phenomenology and enactivism have steered the

[1] I use the word 'phenactivism' to describe views that merge phenomenology and enactivism. Since they share fundamental premises, as will turn out, they can be subsumed under this concept.

research field towards acknowledging previously neglected aspects of social cognition, such as embodiment and interaction.

Presenting these alternative views on social cognition and arguing for my claims, the chapter is structured in three parts. Chapter 2.1. focuses on traditional phenomenological views on intersubjectivity and empathy, such as Scheler’s, Husserl’s, and Merleau-Ponty’s. It will become apparent that they stress the importance not only of the experiential quality of empathy, but also on the role of the body for intersubjectivity. In chapter 2.2., the modern phenomenological perspective is first described and then evaluated. Especially the validity of the central concept of ‘direct perception’ is scrutinized. In the last part, I depict the enactive proposal. In order to understand this view, it will be necessary to first lay out the set of background assumptions of general enactivism, and only then proceed to describe the more specific view on social cognition.

2.1. Traditional Phenomenology

In order to understand the traditional phenomenological conception of empathy and intersubjectivity, it will be helpful to first present some background assumptions held by Husserl and his fellow phenomenologists.

The general goal of phenomenology is expressed in Husserl’s famous attempt to “go back to the things themselves” (Gallagher & Zahavi, 2008, p. 6). What he means is that we are in need of a thorough analysis of the things *as they seem*, as they phenomenally appear to us, focusing on the experiential structure of consciousness and then – using the phenomenological method – bracket the experiences until we get to the ‘things themselves’. Behind this stands the conviction that things in the external world are not different from how they appear to us. Since this is so, investigating conscious experience leads to investigating reality. One essential feature of consciousness, according to Husserl, is intentionality. The notion expresses that consciousness is always ‘about’ something, that it always has an intentional object. These concepts are important for several reasons.

Firstly, Husserl’s goal was to describe empathy as an intentional achievement, i.e., to ask how we experientially grasp the other’s mind. Secondly, when asking what the object of empathy is, we will see that the body plays an extremely important role for empathy and intersubjectivity. By asserting a deep entanglement of mind and body, the Cartesian framework of body and mind as two distinct entities is rejected. Moreover, it is recognized that a plausible specific analysis of empathy needs a plausible general framework of the mind and the body.

In this chapter, I will describe the following aspects of the traditional phenomenological view in more detail:

(1) In chapter 2.1.1., a criticism of previous accounts such as the argument from analogy and Lipps’ conception of empathy are presented. The goal is to show how these critical thoughts build the basis for important phenomenological assumptions, such as the claim that the other can be understood in a *direct* fashion.

(2) I proceed to detail the phenomenological conception of empathy, which is predicated upon the rejection of Cartesian skepticism. Empathy is described as a unique kind of intentionality that comes with its very own perceptual character.

(3) Lastly, the deep relation of embodiment and intersubjectivity is presented. The basic fact of being an embodied agent is seen as a premise for understanding others, since others present themselves as a unity of body and mind. Further consequences of the importance of the body are described and discussed.

2.1.1. Direct Perception and Criticism of Previous Views

Recall that the problem of other minds followed from a Cartesian view of the mind and body. The problem appears because other minds are deemed to be inaccessible without a further inference processes, such as drawing an analogy. In chapter 1.1.2., I have already presented some of Scheler’s criticism of the argument from analogy. Importantly, however, Scheler rejects two more premises of the argument from analogy.

The first states that one’s own consciousness is given to oneself in a *direct* fashion, the second denies such a direct access to the other’s consciousness. Scheler argues that these assumptions on the one hand underestimate the difficulties one can have accessing her own consciousness and on the other hand overestimate the difficulties humans have in accessing the consciousness of another person (Zahavi, 2001, p. 152). The denial of those premises of the argument from analogy are crucial to understand his alternative proposal.

The following quote, which can be seen as the locus classicus of what has been dubbed *direct perception* (e.g., Gallagher, 2008), shows how Scheler (1912/1973, p. 254) vehemently rejects the claim that the only accessible feature of the other is her body, which ultimately means that the other's consciousness is inaccessible for perception and needs an additional mechanism:

> Sicher ist es wohl, daß wir im Lächeln die Freude, in den Tränen das Leid und den Schmerz des anderen, in seinem Erröten seine Scham, in seinen bittenden Händen seine Bitten, in dem zärtlichen Blick seiner Augen seine Liebe, in seinem Zähneknirschen seine Wut, in seiner drohenden Faust sein Drohen, in seinen Wortlauten die Bedeutung dessen, was er meint, usw. direkt zu haben vermeinen. Wer mir sagt, das sei aber keine «Wahrnehmung», da es keine sein «könne», es «könne» aber keine sein, da eine Wahrnehmung nur ein «Komplex sinnlicher Empfindungen» sei und es sicher für Fremdpsychisches keine Empfindung gäbe - und sicher erst keine Reize -, den bitte ich, sich von so fragwürdigen Theorien doch zum phänomenologischen Tatbestand zurückzulenken.

This conception of direct perception implies that the other's mind is not hidden, but is perceivable in her gestures, expressions and behaviour.

Behaviour, gestures and other bodily expressions are, so the phenomenological view, always intentional and hence meaningful, which renders the distinction between the internal and external realm redundant and arbitrary (cf. Zahavi, 2001, p. 153). However, not every phenomenologist agrees with Scheler on his direct perception approach. Husserl, for example, always remained uncertain (Zahavi, 2012) about whether we do have direct access to the other's consciousness or whether this would be a too strong claim:

> Aber nun merke ich, dass das "Seelenleben" des Anderen, dass überhaupt das, was ihn zu einem Menschen und nicht einem blossen Körper für mich macht, bloss "bedeutungsmässig" gegeben ist – "bloss bedeutungsmässig", das ist, keineswegs „eigentlich" wahrgenommen. Nichts vom Psychischen, weder das Psychische im ganzen, die fremde Person, das personale Leben in irgendwelchen Einzelgestalten, irgendein Leiden und Tun, irgendein passives Erscheinendhaben – nichts davon ist in Sonderheit wahrgenommen. Kann Psychisches „wirklich" wahrgenommen werden? Natürlich sage ich, ja. Nur nie das des Andern, vielmehr nur mein eigenes. (Husserl, 1973c, pp. 83–84)

Husserl as well emphasizes that the other presents herself as more than a mere body, but rejects that one can experience foreign psychological states in the same way and as directly as one's own states.

Taking a closer look at Husserl's work reveals that his thoughts on analogy were more ambiguous than Scheler's. Although Husserl does reject the idea of one-directional inference as a means of understanding others, he does not completely banish the notion of analogy. Within his conception of pairing ("Paarung", "Mehrheitsbildung" ibid., p. 15), he claims that my own experience serves as a source of information upon which I draw when transferring meaning to the other's actions. Importantly, this process is fundamentally bi-directional, mutual and does not – like analogy – consist of simple projection.

Up to this point, Scheler and Lipps (see chapter 1.3.1.) do not seem too far apart. However, there has been much criticism of Lipps' positive account of empathy in Scheler's, Husserl's and Stein's work, mostly criticizing him for not providing a phenomenologically valid account of the phenomenon. Thusly, it is criticized that if Lipps' conception of empathy was true, there would be a necessary *transmission* of emotions between individuals. In the case of transmission, though, the own emotional state would collapse as soon as one perceives emotions of another person. This is, however, not the case. As Scheler describes, I may well be sad while still understanding that you are happy.

Moreover, Scheler claims that if the ability to empathize relied upon motor mimicry, I would only be able to understand emotional states that I have experienced myself at some point.[2] Again, this would leave humans unable to understand non-human animals or situations in which one has not been in before. Closely related to this point is Stein's (1917/2008) critique that Lipps' conception does not explain empathy but rather emotional contagion, as is summarized by Zahavi (2010, p. 291):

> In empathy, the experience you empathically understand remains that of the other. The focus is on the other, and not on yourself, not on how it would be like for you to be in the shoes of the other. That is, the distance between self and other is preserved and upheld. Another distinctive feature of emotional contagion is that it concerns the emotional quality rather than the object of the emotion. You can be infected by cheerfulness or hilarity, without knowing what it is about.

This quotation summarizes the difference between emotional contagion and empathy, emphasizing the crucial assumptions held by phenomenologists about these phenomena. Stein concludes that on Lipps' account, there is a mismatch between what he attempts to analyze (empathy) and what he actually analyzes (emotional contagion).

2.1.2. The Phenomenological Conception of Empathy

If the only mind I ever have direct access to is my own mind, nothing else but its contents can serve as a basis for knowing what is going on in the other mind. This Cartesian view is

[2] Interestingly, current research on the mirror neuron system shows that this may actually be the case, at least to a certain degree. In a study with experienced dancers and novices, Calvo-Merino and colleagues (2004) showed that when watching dance videos, activity in the mirror neuron system was stronger in experienced dancers. Thus, if one assumed that the mirror neuron system underlies interpersonal understanding, it could be concluded that my own experiences changes the degree to which I understand another person.

criticized profoundly in the phenomenological tradition. A shift in fundamental background assumptions follows; and some questions simply become redundant. To see how, consider that in the Cartesian framework, there is no other possibility to know of other minds than some kind of inference. In the phenomenological framework, however, the radical Cartesian skepticism is rejected. For if we investigate our perception and experience, we also investigate reality. Note how this is a fundamentally different assumption about the nature and meaning of experience and reality.

Before going into more detail about the phenomenological conception of empathy, let me try to clarify a few terminological issues. There are some confusing manners in how the terms have been used by different authors. The presumably most ambiguous use of terms can be found in Scheler’s work. As Zahavi (cf. 2010, p. 289) shows, he used a variety of terms – *Nachfühlen, Nacherleben, Verstehen, Fremdwahrnehmung* – in addition to empathy. Still, what he most likely meant by all of them can be captured with the concept of empathy. Husserl was more consistent in his usage of words, although he used the term *Fremdwahrnehmung* instead of empathy more frequently in his later writings (cf. ibid.). Edith Stein, whose doctoral dissertation is dedicated to empathy, is more explicit and strict when it comes to empathy. For her, the phenomenon is a unique and irreducible kind of intentionality that is directed at others, and she dismisses any other possible meaning (cf. ibid.). This characterization of empathy as a *special kind of intentionality* is worth a closer look, since it can be assumed that Scheler as well as Husserl would have agreed with that statement.

In order to understand this characterization of empathy, let us consider Husserl’s notion of intentionality. Husserl was a student of and influenced by Franz Brentano, who re-introduced the term intentionality as it is being used until today in the debate of philosophy of mind. It is important to emphasize that intentionality is not equivalent to motivation or intention, but refers to a feature of consciousness. What Brentano and Husserl want to express with the term is that conscious states are always *about* something, i.e., they are directed at something and have a certain content (Siewert, 2011). Husserl’s attempt was not only to investigate intentionality as a fundamental property of consciousness, but also to investigate empathy as an intentional achievement. Accordingly, he describes the relation between intentionality and empathy as follows: “Die Intentionalität im eigenen Ich, die in das fremde Ich hineinführt, ist die sogenannte Einfühlung.“ (Husserl, 1973c, p. 15)

Furthermore, Husserl describes three kinds of intentionality that are arranged hierarchically with respect to how they describe the intentionality relation of an object. The most indirect way is described as *signitive*, referring to mere descriptions of an object, such as me describing to you what my cat looks like when she is lying on the sofa. The *pictorial* way describes instances in which one looks at representations of an object, e.g., a picture of my cat lying on the sofa. The third and most direct way is dubbed *perceptual*, and refers to cases in which one perceives the object itself, e.g., I stand in my living room and see my cat lying on the sofa (Zahavi, 2012).

To which of these kinds of intentionality does Husserl count empathy? There is no straightforward answer to this question in Husserl's work,[3] since he describes the experience of others as *quasi-perceptual* (Husserl, 1984, p. 41), which means that the other is indeed perceived, but not in the same way as one perceives one's own experiential landscape.

Stein and Scheler agree with Husserl that empathy has its own perceptual character, because it is different from the way one perceives oneself, but still is a kind of perception. The empathic process is one that involves apperception[4] rather than inference. To understand what it means that apperception includes 'analogical transference' (Husserl, 1973c, p. 15) but no *analogy*, we must go back to the notion of pairing and how it relates to Husserl's general account of intentionality.

In this general conception, he states that past experiences shape patterns of understanding, namely by analogical transference. Whenever I encounter something again, my prior experience is transferred to the new situation and serves as a repertoire of knowledge on whose basis I perceive and understand the situation. In social encounters, my prior self-experience serves as the basis on which one assigns meaning to the behavior of the other. This is a *mutual* process, which then can be identified as the distinctive feature between analogical transference and analogy. The way Husserl characterizes the former, i.e., as a *bi-*

[3] Zahavi (2012) points out that not only the detailed description of intersubjectivity, but also Husserl's way of describing empathy as an either direct or indirect case of perception has been something the founder of phenomenology has been struggling with throughout his career and never found an definite answer to.

[4] 'Apperception' as it is used by Husserl describes the cognitive act of transforming pure sense data into meaningful perceptions. To clarify, imagine the case of perceiving a three-dimensional object, for example a puppet. Although you cannot see the backside of the puppet, you know that it is there, it is in some way perceived as well. This implicit perception of hidden sides of the object plus the actually perceived sides of it, add up to the whole sensation.

directional process, is fundamentally different from the latter, which entails only a *uni-directional* way of inference.

What we perceive empathically truly is the other and no analogue of my own experiences. Empathy is no reduplication of my own emotional states, but the (quasi-)perception of the other’s experiences. Thus, while analogical inference only gives us access to our own states and leaves us with perceiving some kind of analogue, Husserl’s conception of analogical transference is thought to lead us to truly perceiving the other.

Although Husserl claims that the asymmetry between self and other is not only important, but *constitutive* for empathy, he also states that empathy is more than mere replication of my own emotional states. Therefore, emotional contagion, motor mimicry and imitation do not belong to the realm of empathy (cf. Zahavi, 2010, p. 291).

So far, empathy has been characterized as having a unique perceptual character that is fundamentally different from experiencing one’s own mind. Moreover, empathy is an intentional achievement with which we apperceive the experience of the other in a quasi-perceptual way. Let us now try to grasp the bigger picture and see how this notion relates to the more general issues of embodiment and intersubjectivity.

2.1.3. Embodiment and Intersubjectivity

When empathy is depicted as a unique kind of intentionality, the following question arises: What is the intentional object? Scheler, as well as Husserl, Stein and Merleau-Ponty describe the other – and thus the intentional object of empathy – as a unified whole that presents itself as neither pure body nor pure soul, ultimately rejecting the Cartesian dualistic conception of body and mind as two distinct entities. The question of the intentional object reveals the importance of the body for phenomenological theories of intersubjectivity and empathy.

We have already seen that the notion of pairing implies that self-experience is a precondition for other-experience. This self-experience now boils down to bodily experience, i.e., the experience of myself as a lived body enables the experience of the other as a lived body, which in turn builds the precondition for empathy (cf. Zahavi, 2012, p. 241). Thus, the perception of the other presupposes an understanding of the other’s *body*. This becomes clear in the following quotation in which Husserl emphasizes the importance of the body:

> Erst muss der fremde Leib, und als Zentrum der fremden orientierten Umwelt, für mich da sein, damit sich in ihm etwas ausdrücken kann. In einem blossen Ding kann sich nichts

> ausdrücken, sondern nur <in> einem Ding, dem ein „Leib“ eingelegt ist und in dem in diesem Verstehen sich ein Ichliches weiter indiziert. (Husserl, 1973a, p. 436)

Another crucial point can be found in the quotation: The other is always seen as someone who has a perspective on the world. Experiencing others means to see subjects who experience a world. Thus, the other is not only given as another body, but also as a person with a perspective on the world.

Here lies another reason why empathy can be called a unique kind of perception, it always involves co-experiencing the perspective of the other. This is an interesting thought for the current debate about the question whether social cognition has a set of features that makes it special in comparison to general cognition. Social cognition could be special because the perception of the other always changes my own perception. Cognition that is directed at other individuals thus would not only be shaped by internal processes, but always be influenced by the other.[5]

It is interesting to see how Husserl emphasizes the importance of the first person perspective and the impossibility to ever have perception that does not originate from this perspective. At the same time he stresses – especially within the notion of pairing – how the perception of others profoundly influences my own perspective.

This leads us to the phenomenological attempt to understand the role of empathy in all perceptual objects (Zahavi, 2001) and also to the notion of intersubjectivity.[6] Husserl, as well as Heidegger and Merleau-Ponty, asserts that we are right from the start thrown into a social world and surrounded by others. Heidegger emphasizes that we are not only accompanied by other human beings in a literal sense, but that within all the cultural objects that surround us we always find “indeterminate others” (ibid., p. 154). For him, *Dasein* (being there) always includes *Mitsein* (being with): “Thus, Heidegger ultimately claims that Dasein’s being-with, its fundamental social nature, is the formal condition of possibility for any concrete experience of and encounter with others.” (ibid.)

Although I do not want to go into further detail of Heidegger’s work, his idea nicely shows the deep connection between subjectivity and intersubjectivity, between individual and other that is depicted in phenomenological ideas. For Husserl, perception is intrinsically

[5] A look at studies of perspective taking (e.g., Becchio et al., 2013; Furlanetto et al., 2013) reveals that humans indeed integrate the perspective of another human being under certain circumstances.

[6] See Zahavi (2001) for four different ways in which empathy and intersubjectivity can be or have been related in the phenomenological tradition.

intersubjective in the way that the apperception of objects always includes the perspective of the other. Thus, the perspective of the other, grasped empathically, influences my own perspective and perception.

One of Husserl’s and Merleau-Ponty’s goals was to examine the conditions of possibility for interpersonal relations:

> But this is precisely the question: how can the word ‘I’ be put into the plural, how can a general idea of the *I* be formed, how can I speak of an *I* other than my own, how can I know that there are other *I*’s, how can consciousness which, by its nature, and as self-knowledge, is in the mode of the *I*, be grasped in the mode of Thou, and through this, in the world of the ‘One’? (Merleau-Ponty 1945/2002, p. 406)

Both phenomenologists come to the conclusion that intersubjectivity is possible because of the embodied structure of subjectivity: “Die Einfühlung setzt Leiblichkeit voraus.” (Husserl, 1973b, p. 336).

What exactly does that mean? The most basic assumption is that humans are embodied beings. Husserl refines this general statement and asserts that bodies are two-sided, they entail interiority as well as exteriority (“Innen- und Außenleiblichkeit” ibid., p. 337). Zahavi (cf. 2001, p. 161) clarifies these notions by the example of touching one’s own hand. Not only do I feel my right hand touching my left hand, but at the same time, my left hand feels the touch of the right hand. It is in this double nature of being able to feel the touching and being touched at the same time in the same body that enables intersubjectivity. For Merleau-Ponty (1945/2002, p.434), subjectivity must entail a dimension of the other, or intersubjectivity would not be possible at all:

> How in the first place could I ever recognize other (my)selves? If the sole experience of the subject is the one which I gain by coinciding with it, if the mind, by definition, eludes ‘the outside spectator’ and can be recognized only from within, my cogito is necessarily unique, and cannot be ‘shared in’ by another. Perhaps we can say that it is ‘transferable’ to others. But then how could such a transfer ever be brought about? What spectacle can ever validly induce me to posit outside myself that mode of existence the whole significance of which demands that it be grasped from within? Unless I learn within myself to recognize the junction of the for itself and the in itself, none of those mechanisms called other bodies will ever be able to come to life; **unless I have an exterior others have no interior.** The pluralityofconsciousness is impossible if I have an absolute consciousness of myself. [emphasis added]

The relation of embodiment and intersubjectivity is quite straightforward in this view; without the former, the latter is impossible, being embodied is the condition of possibility of intersubjectivity. Note that this implies a rejection of a dualistic framework, or at least shows that the Cartesian conception must end in solipsism; a disembodied soul could never recognize other persons, let alone their emotional or mental states.

This once more depicts a shift in interrogation. The problem of other minds, with which Lipps was still concerned, simply dissolves in the above described framework. The epistemological question of how disembodied minds recognize other minds is substituted by the more fundamental question of how intersubjectivity is made possible. Assuming that the body functions as a common denominator between subjects and that the ability to understand others is inherent within this incarnated structure, there is no need to ask the epistemological question anymore.

Let me summarize the most crucial points and claims that I introduced in this section about traditional phenomenology. I first reviewed their critique against the argument from analogy. Scheler brought up several critical points and showed that analogical inference as depicted in the argument never leads to knowledge about foreign, but only one’s own mind. Although Husserl and Scheler appreciated Lipps’ criticism against the argument from analogy, they in turn criticize Lipps’ own account of empathy as falling short of actually giving a proper description of empathy. As Stein emphasized, Lipps rather offers an analysis of emotional contagion. The phenomenologists characterize empathy as an irreducible kind of intentionality that is directed towards the experiences of the other, which are then perceived (more or less) directly. Empathy is possible because humans are embodied beings and the structure of the body provides a common basis. Thus, embodiment is constitutive for intersubjectivity.

2.2. The Modern Phenomenological Perspective

The goal of this chapter is to examine what modern phenomenological approaches to social cognition can contribute to the debate. I will argue that while they can help to inform descriptions of the target phenomenon and the experiential nature of social encounters, they do not meet their own goal of offering an account of social cognition which is not only more comprehensive than mindreading approaches, but also empirically plausible.

(a) In subchapter 2.2.1., I will describe the modern phenomenological perspective and its criticism of mindreading accounts of social cognition. The alternative phenomenological proposal and its main three components – primary intersubjectivity, secondary intersubjectivity and the Narrative Practice Hypothesis (NPH) are introduced.

(b) The next section (2.2.2.) examines two conceptions that form the bedrock of the above mentioned proposal, namely direct and smart perception. In an attempt to yield a non- and anti-representational view on social cognition, it is rejected that an 'additional' mental effort such as simulation or theorizing is needed to understand others. Instead it is claimed that perception already provides the necessary means.

(c) Finally, I proceed to evaluate the phenomenological perspective in chapter 2.2.3. and review already existing criticism on this account. I argue that while phenomenology can indeed contribute valuable insights at the phenomenological level of description, its aim to offer a full-blown theory on social cognition founders on its attempt to re-interpret neuroscientific findings.

2.2.1. Intersubjectivity in the Modern Phenomenological Perspective

Most of the background assumptions of modern phenomenology go back to traditional phenomenologists such as Scheler, Husserl, Gurwitsch, Merleau-Ponty, but also to the work of psychologists like Gibson, Trevarthen, and Meltzoff (Gallagher, 2001, 2005, 2008; Zahavi, 2001, 2010, 2011, 2012). Gallagher (cf. 2005, p. 95) stresses that his views are *inspired by* Husserl or *Husserlian*, but that it is yet an open question whether they are fully compatible with Husserl's view on intersubjectivity, i.a., because it is unclear whether Husserl ever settled with his own theory on the topic.

In one of his works, Gallagher describes parts of Husserl's, Scheler's and Gurwitsch's accounts and aims to show that these views of intersubjectivity are compatible with neuroscientific findings. The views of these phenomenologists are, so he claims, more than a phenomenological description of intersubjectivity, they offer a full-blown theory of social cognition (Gallagher, 2005). He concludes that

> [w]hen taken together, these two analyses go a long way toward providing the best explanation of the most basic processes of intersubjectivity, and they provide an account that is superior to any of the standard accounts in circulation today, for example theory theory or simulation theory. […] And finally, as mentioned, these phenomenological accounts are fully supported by recent discoveries in the neurosciences that show the importance of mirror neurons, and neural representations that are activated both when I engage in intentional action and when I observe someone else acting. (ibid., p. 106)

Gallagher’s (2001) paper *The Practice of Mind: Theory, Simulation or Primary Intersubjectivity?* is often described as an important point in the development of theories of social cognition and as one of the first publications that offered a profound criticism of the then-prevalent ST/TT paradigm. Gallagher not only doubts assumptions that are at the very core of ST and TT, but also offers an alternative theory that is supposed to remedy the shortcomings of mindreading accounts. The basic claims of the paper have been extended and refined in many follow-ups, but have not succeeded to replace ST and/or TT.

Since one attempt of proponents of modern phenomenological approaches is to make claims that are along the lines of findings and results in cognitive science, their view is often called ‘neurophenomenology’ (e.g., Thompson, 2010). Importantly, they reject some basic assumptions of ‘mainstream’ neuroscience, such as representationalism, reductionism and mechanistic explanations (Rowlands, 2009) and follow the enactive account of cognition. The following quotation of two leading proponents of phenomenology sums up their focus:

> Phenomenlogical views, then, involve non-mentalizing, embodied, perceptual approaches to questions of understanding others and the problem of intersubjectivity. We begin from the recognition that the body of the other presents itself as radically different from any other physical entity, and accordingly, that our perception of the other’s bodily presence as a *lived body,* a body that is actively engaged in the world. […], it would be a decisive mistake to think that my ordinary encounter with the body of another is an encounter with the kind of body described by physiology. The body of another is always given to me in a situation or meaningful context, which is co-determined by the action and expression of that very body […].” (Gallagher & Zahavi, 2008, p. 183)

This shows a clear emphasis on the body, which has already been put forth by Husserl and Merleau-Ponty. The traditional view that intersubjectivity is an irreducible kind of empathy is picked up and specified to the claim that it is irreducible to *mindreading processes* that are internal and mostly exclude the non-neural body.

Note that extra-neural structures of *both participants* of a social encounter are of importance in this view. The other’s behavior is thought to be expressive, it is “already from the start soaked with mindedness” (Zahavi, 2011, p. 550) and thus the other person’s mental states can be perceived directly.[7] At the same time, the perceived behavior elicits a reverberation in one’s own body which then enables understanding the other:

[7] Zahavi (2011) sorts out different notions of ‘directness‘ that cause confusion. While Jacob (2011), seems to criticize a notion of directness with the claim that perception relies on contextual cues and thus should not be described as direct, Zahavi (cf. 2011, p. 548) states that the phenomenological

> When we see someone else act in a certain way, our own kinaesthetic system is activated in a way that mirrors the perceived action. This, in part, is what allows us to understand the other person. Moreover, and importantly, this kinaesthetic activation is part of the perceptual process – part of the hyletic processes that underpin the *noetic* aspect of perception. (Gallagher, 2005, p. 97)

The quotation not only shows how Gallagher aims to couch his Husserlian theory in terms that are compatible with current scientific findings (mirroring and mirror neurons), but also that the conception of perception is fundamentally different from most cognitivist views in that it counts kinaesthetic processing as belonging to the perceptual domain. Based upon this background, Gallagher and colleagues have developed an account of social cognition that is thought to offer developmentally and empirically plausible conceptions. This account consists of three further notions, namely primary intersubjectivity, secondary intersubjectivity and narrative practice.

Primary intersubjectivity is thought to be primary in a double sense; not only is it thought to be the ability that enables newborns and infants to meaningfully interact with their caregivers, it also stays the prevalent way of understanding others throughout life (Gallagher, 2001). This kind of intersubjectivity is a non-mentalizing skill which is constituted by embodied practices and thus neither needs simulation nor theorizing.

Since mindreading capacities are claimed to arise rather late in life (e.g., the false belief task predicts ToM to develop around the age of four), or respective theories stay quiet about when children acquire their crucial capacities, the notion of primary intersubjectivity is seen as giving a more plausible account from an ontogenetic perspective. Primary intersubjectivity neither requires the postulation of inference of anything hidden, nor does it need additional sophisticated cognitive processes – it is enabled via direct perception which is the only mechanism required to grasp the other’s meaningful expressive behavior (Gallagher & Hutto, 2008).

Social interactions are always embedded in a certain context and situation. Taking these into account when interacting and tying actions to pragmatic contexts is central to what is called secondary intersubjectivity, which develops around the age of one (ibid.). Because the situational and environmental context of a social interaction provides informational contents

usage of the word stands in contrast with ‘mediated’ and thus only rejects *cognitive* additional acts, not context-dependency.

which then serve as a "shared motivational background", it is not necessary to bother reading the other person's mind (cf. Gallagher & Zahavi, 2008, p. 191).

However, while phenomenologist are convinced that these embodied practices are sufficient for understanding most social situations, they admit that sometimes there is more to grasping the reasons for other's behavior (Gallagher & Hutto, 2008). The Narrative Practice Hypothesis (NPH) describes the third capacity a child needs in order to make sense of other people. The child learns, so it is claimed, about reasons of actions and normative rules by the pervasiveness of narratives that she is surrounded by. The stories that children are being told – something that seems to happen often enough in every child's life, according to Hutto (2008) – teach them norms and stereotypical behavior. The acquisition of folk psychological knowledge thus depends crucially on story-telling.

2.2.2. Perception: Direct and Smart

The bedrock of the modern phenomenologist conception of social cognition is a specific notion of perception. In stark contrast to representational theories of perception, it is assumed that perceptual processes are not peripheral to cognition, but are much richer and do not entail an additional cognitive step. Perception, according to these views, is direct and smart. Let me try to make sense of this claim in the following.

What does it mean that something can be *directly* seen? We have already encountered Scheler's account of direct perception, but let me introduce another influential view that is of importance. Gibson (1979) introduced the concept of direct perception which relates to his famous conception of *affordances*: "The affordances of things for an observer are specified in stimulus information. They seem to be perceived directly because they are perceived directly." (Gibson, 1977, p. 79) Affordances are directly perceivable since they are physically real, which means that they exist independently of the perceiver. As such, they are perceivable properties of the environment (cf. Gibson, 1979, p. 129). This view is crucially different from the representationalist assumption that object properties need to be mentally represented and perception thus requires an intermediary step.[8]

[8] In the following, I will use the requirement of intermediary steps as the distinctive feature that differentiates directness and indirectness. In doing so, I follow De Vignemont (2010, p. 291): "There is a direct access if and only if the causal transmission of information is direct and does not involve intermediary steps."

Although Gibson is probably the most famous proponent of direct perception, Gallagher (cf. 2008, p. 532) explains in a footnote that his perspective is not to be entirely equated with this Gibsonian view. He emphasized that he does not deny the sub-personal complexity of perception. Much rather, he counts these complex mechanisms as part of the perceptual process. To capture this difference, the conception of *smart perception* is introduced:

> But this informing process is already built into the perceptual process so that as I consciously perceive, my perception is already informed by the relevant sub-personal processing. I don’t first perceive and then add memory in order to recognize my car. My perception, in this sense, is direct even if the sub-personal sensory processing that underpins it follows a complex and dynamic route. (ibid., p. 537)

Even with that kind of definition, his view still presupposes that there are properties of external objects that can be ‘directly’ picked up, that exist independently from the perceiving subject. As such, it is indeed *reminiscent* of a Gibsonian conception.

With the idea of direct perception, the need for representation vanishes. It thus motivates an anti- and non-representational view:

> Rather than saying that I represent my car as drivable, it is better to say that – given the design of the car, the shape of my body and it’s [sic] action possibilities, and the state of the environment – the car is drivable and I perceive it as such. (Gallagher & Zahavi, 2008, p. 8)

Note how this implies a fundamental criticism of mindreading theories, which assume that an ‘additional’ process like simulation or theorizing is needed to understand others. According to direct perception proponents, all necessary processes are already inherent in perception, no further cognitive steps are needed. If the external world and the minds of others, as assumed by Gallagher and colleagues, can be perceived *directly*, however, no such inference is needed:

> The mental states of others are not hidden, and need not to be inferred on the basis of perceiving the behavior; rather, behavior is an expression of the mental phenomena that, in seeing the behavior, is also directly seen. (Newen, 2015, p. 5)

This is why Gallagher (2008, p. 535) describes direct perception also as “smart perception” in contrast to “not-so-smart-perception” (i.e., perception that only serves as input for processes like simulation or theorizing that are thought to do the dirty work and carry the cognitive burden):

> The smarter the perception is, the more work it does; the dumber it is, the more it requires extra cognitive processes (theory, simulation) to get the job done. The direct perception theorist is claiming that social perception is very smart and that in the usual circumstances

> of social interaction it does most of the work without the need of extra cognitive (theoretical or simulationist) processes. (Gallagherm 2008, p. 538)

Perception is described as smart and direct because it is *informed by* many processes. Importantly, these can consist of neural mechanisms (e.g., mirror neuron mechanisms) and extra-neural (e.g., bodily) routes. The latter also include objects that offer affordances – my car is perceived *as driveable* (cf. ibid., p. 537), for example. Perception is also informed by emotional processes that then bias and influence how an object or a person is perceived (cf. ibid., p. 538).

To briefly recap, direct perception is characterized as a mechanism that is informed by sub-personal processes, making it rich and smart. ‘Direct’ thus means unmediated, since no *additional* cognitive act is needed in order to gather the relevant knowledge for understanding others.

2.2.3. Evaluating the Phenomenological Perspective

The main argument for direct and smart perception draws on phenomenal experience and states that since neither simulation nor theory show up in experience frequently, they cannot be ubiquitous capacities that our social understanding depends on (Gallagher & Zahavi, 2008). Behind this stands the claim that the experienced smoothness and immediacy of social encounters has an epistemic value in the sense that it tells us something about the access of other minds. “Directness”, however, is a concept that is used in academic research that is relative to a specific level of description. What does that mean for the validity of the phenomenological view?

To see this, we have to scrutinize Gallagher’s standpoint that smart perception is a sub-personally enriched and informed mechanism, which enables understanding others without ‘additional’ mental effort. This view is largely dependent upon a re-interpretation of neuroscientific evidence of the mirror neuron system, usually interpreted as underlying mechanisms of simulation. Gallagher (2008, p. 541) contends that the rapid activation of mirror neurons (30-100ms after activation of the visual cortex) makes a distinction between perceptual and additional processes redundant:

> A distinction at the neural level between activation of the visual cortex and activation of the pre-motor cortex does not mean that this constitutes a distinction between processes that are purely perceptual and processes that involve something more than perception

It seems to me that one important question follows from this interpretation of empirical data, namely how to individuate mental processes and mechanisms. Gallagher's individuation criterion here is *temporal proximity*, which appears rather arbitrary. I find is questionable to claim that temporal correlation justifies the assumption of mechanistic inseparability. Perception – in Gallagher's view – could entail almost any process as long as it appears in a more or less specific time window.

As an alternative and more reliable criterion, I propose to individuate mental mechanisms on the basis of their *functional properties*. In contrast to temporal properties, they are more substantial and conceptually relevant features and furthermore enable a more fine-grained view on sub-personal mechanisms underlying social cognitive processing. If mechanisms are individuated by their functional role instead of the temporal properties of the physical realizers of this function, however, perceptual and mirror neuron mechanisms are distinct. If this is true, it is unfeasible to keep up the concept of 'smart perception', since this would presuppose that perceptual and post-perceptual processes can actually be described as *one* mechanism. 'Direct perception' as a term which describes the core of understanding others also loses its significance, because mirror mechanisms are seen as functionally distinct, forming an intermediary and thus 'additional' step in the process of making sense of another person.

We can now come back to my main point, namely that directness is relative to the level of description. At the phenomenological level of description, the concept of direct perception may apply in the sense that we do experience ourselves to immediately get a grasp of the other person. However, this experiential quality of directness has complex and indirect sub-personal counterparts. The sheer closeness on our everyday timescale does not justify anything on other levels of description, which is why I deem Gallagher's notion of smart and direct perception incoherent as a description at the sub-personal level. Direct perception, or so I argue, should be treated as a phenomenal quality and thus should not be mistaken with the epistemic mechanism itself.[9]

Another weakness of the theory of direct ad smart perception seems to be that *parsimony* is taken to be a feature of cognitive processes. It is argued that there is 'no need' to describe

[9] This has been described by Metzinger and Windt (2015, p. 7) as the "*E-error*: a category mistake in which epistemic properties are ascribed to something that does not intrinsically possess them." See also Metzinger & Windt 2014.

social cognition as a process which entails several distinct mechanisms. However, parsimony is a property of *theories* of cognitive processes and not of those processes themselves.

Let me now address the possible explanatory scope for social cognition of the phenomenological perspective. It is somewhat unclear whether proponents of phenomenology agree on what their paradigms can offer and what is out of their scope. While Gallagher (cf. 2005, p. 106) often does not hesitate to state that phenomenological theories can *explain* intersubjectivity and feels comfortable to interpret scientific findings on sub-personal processes that have been related to social cognition, he and Zahavi (2008, p. 177) at the same time claim that “phenomenology doesn’t give us access to the sub-personal domain”.

Given this criticism I just offered, it can be questioned whether the phenomenological perspective can really offer an alternative view on social cognition and if mindreading approaches should thus be fully rejected. Although it is conceded that in some cases mindreading abilities need to kick in to understand another person, phenomenologists insist that these incidents are extremely rare. I hence take it that the goal striven for is to provide an alternative to mindreading accounts.

This can also be found when looking at how Gallagher addresses a potential objection to his claims; namely that perception could never be so smart as to reach hidden mental states like intentions. He counters that the very idea of necessary inference because of hiddenness is an ill-conceived notion to begin with. Social encounters do not, according to phenomenologists, consist of observation and inference, but of interactions that happen to be embedded in a specific context (Gallagher, 2008; Zahavi, 2011). This rejects the theoretical core of ST and TT and proposes a different foundation of social cognition: interaction.

Interaction is thought to offer perceptual cues that are responded to in an embodied manner while contextual information helps to disambiguate stimuli. In Gallagher’s (2008, p. 540) words:

> What we call social cognition is often nothing more than that social interaction. What I perceive in these cases does not constitute something short of understanding. Rather my understanding of the other person is constituted within the perception–action loops that define the various things that I am doing with or in response to others.

This quotation entails several important claims of the phenomenological perspective. It firstly rejects the way mindreading paradigms have tried to account for social cognition, namely in ways that neglected interaction. The alternative description given here focuses on action and interaction instead of mindreading, thus replacing the *explanandum* for social cognition research. Behind that stands the claim that there is little phenomenological evidence that we infer hidden mental states as ubiquitously as stated by ST and TT. This leads to the second claim, namely that the *explanantia* that have been offered within paradigms that focus on mindreading instead of interaction need to be substituted. As Slors (2012) emphasizes, while phenomenologists may be right to reject that inference is ubiquitous at the *personal* level, they have no basis for rejecting a notion of sub-personal inferential processes. In combination with the criticism I lined out above, it seems that the phenomenological perspective struggles to offer an account that would be able to fully substitute ST and TT.

Overgaard and Michael (2013) put forth a similar view. However, while they come to the conclusion that direct perception fails to reach the goal of being an alternative to ST and TT and thus should be considered as one out of many processes that enable social understanding, I argue for a rejection of direct perception as a mechanism that enables social cognition. Direct perception can describe the experiential nature of (some) social encounters at the phenomenological level, but it cannot, as argued above, be taken seriously as a cognitive mechanism. I thus disagree with Overgaard and Michael’s claim that direct perception can describe a functional mechanism of social cognition.

A theory of social cognition that aims to provide a comprehensive picture of the phenomenon needs to be able to account for empirical results in a plausible manner. I have suggested that ‘interdisciplinarity’ should be among central desiderata of such a theory, which makes it even more important to interpret experimental work in a non-arbitrary way. I thus introduce the next desideratum that is supposed to capture this aim:

d_5 – empirical plausibility
Empirical plausibility demands that the empirical predictions a theory makes are consistent with actual empirical findings.

At the same time, the contribution that phenomenological accounts have made to the debate should not be underestimated here. Taking into account the phenomenal character of social encounters and thus being able to yield in-depth descriptions at this level of analysis is an important point that has only gained attention after the modern phenomenological view has been worked out in more detail.

This critical appraisal shows that phenomenological approaches can help improving the description of social cognition as the target phenomenon. By pointing to the fact that there is more to understanding others than consciously thinking about them, Gallagher and his colleagues have made a precious contribution to the debate. However, since the description of phenomenal experience is only one level that needs to be considered for a full-blown theory of social cognition, I propose to let phenomenological theories not count as candidates for this exercise.

2.3. Making Sense of One Another – Enactivism

The point of departure of enactive approaches to social cognition is the claim that neither mindreading nor phenomenological theories are able to provide an exhaustive account of the phenomenon. Enactive accounts aspire to thoroughly substitute mindreading approaches and, in doing so, to get rid of a brain-bound cognitivist picture of the mind. In this chapter, I will outline the claims and arguments made by proponents of an enactive approach and show that they are only valid if one accepts their metaphysical background assumptions. In doing so, I proceed as in the following order:

(1) In chapter 2.3.1., I present the general enactive account of the mind as the background theory. This extensive review is deemed necessary in order to understand that the enactive view cannot be taken to simply add a little bit of action and embodiment to the debate, but forms a proper alternative to cognitivist views of the mind.

(2) Based upon this background is the notion of participatory sense-making, which forms the enactive alternative to social cognition. In chapter 2.3.2., the principles and core claims will be presented.

(3) By formulating the Interactive Brain Hypothesis (IBH), proponents of the enactive view broaden the explanatory scope and aim to show that their view includes a plausible account of the brain. The claim here is that all mechanisms of the social brain depend on interaction – either developmentally or contemporarily – even in the absence of interactive situations. I will describe this view in chapter 2.3.3.

(4) Chapter 2.3.4. depicts the perceptual crossing paradigm (PCP) as the alleged main empirical evidence for the enactive view and discusses several follow-ups.

(5) Finally, I critically assess the enactive perspective with respect to the validity of their main claim, namely that interaction dynamics constitute social cognition. With

respect to this assessment, IBH and PCP are also evaluated. I will argue that it is far from clear whether PCP indeed serves as an empirical back-up for IBH and the claim that interaction constitutes social cognition.

2.3.1. The Embodied Mind

In the early 1990s, Varela, Thompson and Rosch (1993) published their book *The Embodied Mind* in which they aimed to provide a non-cognitivist, alternative model of the mind. Their motivation was to criticize the computer model of the mind which they claimed to be unsatisfactory since it lacks a pragmatic approach to cognition and consciousness and misses to integrate the inherent connection between mind and life (Thompson, 2010).

In this section, I will depict the enactive view of the mind which was first formulated by the authors named above but is now to be found in a variety of versions and subversions. Since the most famous enactive approaches to social cognition that have been put forth in the beginning of the 21st century are based on this 'original' version of enactivism, I will focus on claims along those lines. Note that I will not restrict myself to the claims made in the monograph (The Embodied Mind), but also include refinements of the theory that have been developed after the timespan this chapter scrutinizes. However, in this subchapter I will not portray the enactive view of *social* cognition, but lay the foundation to do so by describing the central ideas and background assumptions of enactivism here.

Enactivism is couched in dynamical systems theory (DST) and seconds its basic claims. Without going into too much detail of DST, there are several aspects which are important to mention here because they exhibit fundamental differences to cognitivist claims. Thompson describes the most important divergence to be the view on cognitive systems in general. While DST states "that natural cognitive agents (people and other animals) are dynamic systems (or, more precisely, that the cognitive systems agents instantiate are dynamic systems), and that accordingly action, perception, and cognition should be explained in dynamic terms" (Thompson, 2010, pp. 40–41), cognitivism claims that "cognitive agents (or the cognitive systems they instantiate), whether natural or artificial, are digital computers or physical symbol systems and that accordingly cognition should be explained in symbol-processing terms." (ibid., p. 41) Thus, the *unity of analysis* is different in the two approaches to cognition. As will turn out in this section, this fundamental difference entails consequences that leave the theories hard to reconcile.

In general, DST aims to model changes of a system’s states over a certain timespan (hence, dynamic) and to thereby predict and explain its behavior. A system possesses a state space which describes the space of all potential states the system can be in. What a model of these systems tries to capture is the trajectory of state changes which can emerge spontaneously in virtue of the system’s self-organizing structure (cf. Thompson, 2010, p. 42). Systems and its patterns are thus dynamic, non-trivially complex,[10] and occur gradually. It is claimed that these properties cannot be assigned if a cognitivist view is adopted which depicts cognition in rather static and all-or-nothing terms. The view of cognitive processes as unfolding over time in a spontaneous and dynamic manner, according to Thompson, cannot be captured within the sandwich view of the mind with a predetermined order of processing from input to output (cf. ibid., p. 43).

Proponents of the enactive view not only reject the idea of mental processes as being functional and thus multiply realizable, but also challenge the view of the mind as an input-output device. A ‘radical’ cognitivist picture of the mind is fully internal and neglects the role of the body entirely, at least in some versions. Enactivists adopt the Gibsonian rejection of the distinction between inner and outer, claiming that the first mistake to make in thinking of cognition is to assume that it has a location which is found either inside or outside the skull (Arnau et al., 2014).

Cognition is continuous with biological life and thus so deeply entangled that no distinction would be plausible to defend. This also implies the denial of one core assumption of cognitivism, viz., representationalism. Instead of relying on mental representations that are elicited by (neural) computations, cognition is depicted as *embodied action,* and the mind is seen as an autonomous network that emerges (in a strong sense) within this embodied action. More specifically, enactive views build on five core ideas (cf. De Jaegher & Di Paolo, 2007, pp. 486–488; Thompson, 2010, p. 13):

(a) autonomy,[11]

[10] Thompson (cf. 2010, p.40) states that complexity in terms of DST describes behavior as being neither random nor orderly, but rather as having changeable, and thus unstable patterns.

[11] This is closely related to *autopoiesis*, which is defined as follows by Di Paolo and Thompson (2014, p. 69): “The concept of autopoiesis describes a peculiar aspect of the organization of living organisms, namely, that their ongoing processes of material and energetic exchanges with the world, and of internal transformation and metabolizing, relate to each other in such a way that the same organization is constantly regenerated by the activities of the processes themselves, despite whatever variations occur from case to case.”

(b) emergence,
(c) embodiment,
(d) sense-making,
(e) experience

In the following, I will unpack these concepts in greater detail and aim to show that enactivism comes with both ontological as well as methodological commitments. The point of departure for understanding all of these concepts is the idea that cognitive science needs to deal with and scrutinize *whole biological organisms and their environment* in order to understand how cognition works. This follows from the view that cognition is a property that only arises within embodied and situated beings and thus cannot be restricted to the brain.

The first concept refers to the idea that organisms are autonomous agents which maintain and generate their identity, thereby demarcating themselves as distinct (hence autonomous) entities from (but also in exchange with) their environment. In the words of De Jaegher and Di Paolo (2007, p. 487): “An autonomous system is defined as a system composed of several processes that actively generate and sustain an identity under precarious conditions.”

The notion of *precariousness* entails that systems are constantly fighting against extinction and thus need to sustain themselves. Within this process, cognition emerges (cf. Thompson, 2010, p. 13). Note the further implications of this view. It displays the conviction that biological life and cognition are ‘inextricably entangled’ and cannot be separated, both explanatorily (Thompson & Cosmelli, 2013) and ontologically (Varela, Rosch & Thompson, 1993).

This view entails fundamental rejections of some cornerstones of cognitivism. It means that functionalism is rejected, and also speaks for a *non-reductive* naturalism that depicts cognitive systems as not merely responding to external stimuli and satisfying internal demands, but as beings that are *inherently active*. The notions of response and internalism, according to the enactive view, “fail to give the autonomous agent its proper ontological status.” (De Jaegher & Di Paolo, 2007, p. 487) It is important to note that although the centrality of the brain as the sole organ of the mind is rejected, the relevance of the nervous system is still acknowledged. However, it is seen as *being part* of a sensorimotor system. Along those lines, Thompson (2010, p. 47) states that “[i]n all animals, neuronal networks establish and maintain a sensorimotor cycle through which what the animal senses depends directly on how it moves, and how it moves depends on what it senses”, and furthermore that “every animal meets the environment on its own sensorimotor terms.”[12]

The notion of autonomy is closely related to that of emergence. In autonomous agents, what emerges is an entity that brings forth its own coupling with its environment and thus creates an identity. Importantly, the notion of emergence is modified in the enactive approach and re-described as “dynamic co-emergence” (Thompson, 2010, p. 60). This means that there are basically two processes which “co-emerge and mutually specify each other” (ibid.). The first is local-to-global determination, i.e., the emergence of macro-level processes from the collective of behavior components. These newly emerged processes constrain and control micro-level structures, which is captured in the second notion of global-to-local determination. The result is a circular causality since global behaviors constrain the behavior of local components while the behavior of these components generate the global order (cf. ibid., pp. 61–62). The relation between wholes and parts is summarized by Thompson (ibid., p. 65) as follows:

> In an autonomous system, the whole not only arises from the (organizational closure of) the parts, but the parts also arise from the whole. The whole is constituted by the relations of the parts, and the parts are constituted by the relations they bear to one another in the whole. Hence, the parts do not exist in advance, prior to the whole, as independent entities that retain their identity in the whole. Rather, part and whole co-emerge and mutually specify each other.

Furthermore, these processes are described as arising spontaneously and in a self-organizing manner.

This has important implications for the enactive view of the mind and cognition, since it dispenses the theory from assuming a homunculus, or ‘self’ as a control entity. As De Jaegher and Di Paolo (cf. 2007, p. 487) show, taking emergence seriously also means to be skeptical about the possibility of localization of cognitive functions at either low or high levels and ultimately leads to the rejection of a view of the mind as having circumscribed modules with specialized functions.

Thus far, cognition as put forth by enactivists can be described in the words of Froese and Di Paolo (2011, p. 18):

[12] This view is reminiscent of Gibson’s (1979) description of affordances. Gibson claims that perception depends on features that are attributable to the individual, as well as properties of the object.

> Cognition is the regulated sensorimotor coupling between a cognitive agent and its environment, where the regulation is aimed at aspects of the coupling itself so that it constitutes an emergent autonomous organization in the domains of internal and relational dynamics, without destroying in the process the agency of that agent (though the latter's scope can be augmented or reduced).

Autonomy and its close reference to biology relate to the next core concept of embodiment. It is claimed that cognition is necessarily embodied, "it cannot but be embodied" (De Jaegher & Di Paolo, 2007, p. 487). This follows quite trivially if one rejects neurocentrism and accepts the claim that a biological organism as a whole actively brings forth its own cognitive domain (cf. Thompson, 2010, p. 13).

By making these claims, enactivists explicitly refer to the work of Merleau-Ponty and his conception of embodiment, as the following quotation shows: "For Merleau-Ponty, as for us, *embodiment* has this double sense; it encompasses both the body as a lived, experiential structure and the body as the context or milieu of cognitive mechanisms" (Varela, Rosch & Thompson, 1993, xvi). The centrality of the body for cognition also means that accounts of the mind which neglect or diminish the role of bodily processes are fundamentally flawed. Within their embodied activity, agents not only actively regulate their coupling with the environment, they thereby establish a perspective onto the world (cf. De Jaegher & Di Paolo, 2007, p. 488). Agents thus *create* meaning, there is no passive reception of information which is processed into or in virtue of internal representations which then (potentially) bear meaningful content. This idea is captured by the notion of sense-making, which is characterized as "a relational and affect-laden process grounded in biological organization." (ibid.)

It is interesting to look at the status that is assigned to information and meaning created within this process. Information in cognitivist terms is described as being based on an objectivist conception that views representations as encoding context-independent structures about the world. This conception only works, however, in a heteronomous – as opposed to autonomous – view of agents (cf. Thompson, 2010, p. 52). A theory that depicts agents as autonomous states that information is context-dependent, agent-relative and hence is determined by the agent.[13] The distinction between representational and enactive views of

[13] "Information is context-dependent and agent-relative; it belongs to the coupling of a system and its environment. What counts as information is determined by the history, structure, and needs of the system acting in its environment." (Thompson, 2010, pp. 51–52)

cognition thus boils down to be between “autonomous meaning-construction and heteronomous information processing” (Thompson, 2010, p. 54). The thought to keep in mind here is that exchanges between the organism and environment – i.e., physical processes – are seen as *inherently meaningful.*[14]

The last concept of experience carries several important aspects that are central to the enactive approach, since it displays the close proximity to phenomenological theories and how they mutually inform each other. Experience is taken to be no epiphenomenon, but an issue that is central to the endeavor of understanding the mind (cf. ibid., p. 13). Along the lines of the conviction that mind and life are inextricably entangled, experience is seen as arising naturally within the enaction of a world and thus is not seen as the puzzling case. A close look at the enactive conception of experience reveals a set of interesting background assumptions.

Thompson (ibid. p. 15) claims that one of the deep convergences of enactivism and phenomenology manifests itself in the fact that “both share a view of the mind as having to constitute its objects”. To constitute, however, is then defined as to present, to bring to awareness, to disclose, and not as to fabricate or create. There are several questions that arise when scrutinizing this conception. First, does this statement entail an *ontological* or an *epistemological* claim? The second question refers to the notion of information as being determined and constituted by the agent, thus provoking the question of the ontological status of the world.

To constitute something usually expresses a metaphysical or ontological relation in the sense that A exists *in virtue of* B. If this was the sense in which Thompson uses constitution, this would speak for a strong constructivist assumption that denies the observer-independent existence of the external world. The following quotation could be seen to entail such a claim: “[…] a cognitive being’s world is not a prespecified, external realm, represented internally by its brain, but a relational domain enacted or brought forth by that being’s autonomous

[14] However, there are also rather vague claims which depict meaning as something non-physical: “The distinction between a strictly physical encounter and a cognitive one is to be found in the dimension of significance for the cogniser itself that is characteristic only for the latter class, even though cognitive interactions are themselves also physical encounters.” (De Jaegher & Di Paolo, 2007, p. 488). This becomes even more obvious in Fuchs and De Jaegher’s (2009, p. 471) description of coordination of two embodied agents: “It should be clear that, by focusing on the coordination dynamics of the process of interacting, we are not assuming a purely physical process. It is embodied subjects who coordinate, which means that in these couplings there is also a coordination of *meaning.*”

agency and mode of coupling with the environment” (Thompson, 2010, p. 13), or in the words of Fuchs and De Jaegher (2009, p. 470): “a cognitive being’s world is not a pregiven external realm represented by the brain. Rather, it is the result of a ‘dialogue’ between the sense-making activity of an agent and the responses from its environment.” Both quotations can be interpreted in two ways. ‘A cognitive being’s world’ could either refer to the external, physical world that the agent lives in. On this reading, enactivists would make the strong claim that the world only exists in virtue of the sense-making activity of the agent and thus clearly entail an ontological claim. If this ‘world’, however, does not entail more than a being’s *experiential* world, the claim is weaker or at least more ambiguous, since it would only relate to the conscious perception of the world and make no statements about the ontological status of the physical world. The claim would then rather entail an epistemological statement and make assumptions about how cognitive agents discern the world. By re-defining constitution as bringing to awareness, the authors seem to mix up ontological and epistemological claims which makes their position hard to grasp.

While it seems that it took longer for cognitive science to pick up enactivist claims, the hypothesis I will unpack in the next section had an immediate impact on the research field of social cognition.

2.3.2. Participatory Sense-Making

Before unpacking the central concept of participatory sense-making, let me briefly describe the motivation for introducing such a radical alternative view of social cognition. In 2007, De Jaegher and Di Paolo published a seminal paper that aimed to lay out such a view. Importantly, the authors not only reject mindreading accounts, but also claim that the phenomenological perspective has flaws and leaves important aspects unexplained. Still, enactive accounts of social cognition see similar shortcomings of mindreading views as phenomenologists and basically state that simulation and theorizing are not at the center of social abilities.[15] Despite their close proximity to and convergences with phenomenological theories,[16] they have been criticized to be unable to interpret central empirical paradigms and thus fail to yield an account that puts interaction in its center.

[15] Note that this rejects the *centrality* of mindreading for social cognition without denying its *existence*; see Di Paolo & De Jaegher, 2012, p. 2.

[16] For an attempt to frame social cognition in terms of both enactivism and phenomenology, see Fuchs & Jaegher, 2009.

This is exemplified by showing that the phenomenological interpretation of a study which investigates interaction processes between mother and child within the framework of primary intersubjectivity is not sufficient. In Murray's and Trevarthen's (1985) famous double-TV experiment, an infant is seated in a room separate from her mother and is presented with a TV screen on which she either sees a live recording of her mother through which she is able to communicate with her, or a pre-recorded video of the mother interacting with the infant. While in the live-condition the infant engages with the mother and interacts with her, she shows clear signs of distress when presented with a pre-recorded recording of the caregiver. According to the phenomenological direct perception approach, the mere perceptual input of the mother's interactive behavior should be enough to entertain the child. Primary intersubjectivity thus cannot explain the distress of the child, since it makes no predictions about the relevance of timing and coordination and their breakdowns. It was concluded that there must be something crucial *in the interaction itself* that explains the sensitivity of the infant to these subtle cues (cf. De Jaegher & Di Paolo, 2007, p. 490).

The centrality of interaction is the core assumption of enactive accounts and builds the starting point for further claims. Social interactions are seen as providing enabling conditions and forming constitutive elements for both the development and maintenance of social skills (De Jaegher & Di Paolo, 2007; De Jaegher, Di Paolo & Gallagher, 2010; Di Paolo & De Jaegher, 2012). In order to expound this view, the claim is couched in theoretical terms of DST and enactivism. Empirical set-ups, such as the double TV-experiment and most importantly the perceptual crossing paradigm (see section 2.3.4.) are assumed to corroborate these theoretical aims. At this point it will be helpful to look at how proponents of the theory conceive of interaction. Here is a definition that now seems to be widely accepted:

> Social interaction is the **regulated coupling** between at least two autonomous agents, where the regulation is aimed at aspects of the coupling itself so that it **constitutes an emergent autonomous organization** in the domain of relational dynamics, without destroying in the process the autonomy of the agents involved (though the latter's scope can be augmented or reduced). (De Jaegher & Di Paolo, 2007, p. 493) [emphasis added]

This definition shows that interactions are viewed as building autonomous systems which then are metaphysically irreducible to mechanisms of the individual that are involved. Two systems are furthermore said to be coupled when their behavior and mental states depend on each other.

The concept of 'participatory sense-making' has been introduced to capture these ideas. De Jaegher and Di Paolo (ibid., p. 497) define the term as "the coordination of intentional activity in interaction, whereby individual sense-making processes are affected and new domains of social sense-making can be generated that were not available to each individual on her own.” Together with the definition of interaction given above, this means that interacting individuals stand in very close relation to each other. Since sense-making can be seen as the enactive term for cognition, these claims boil down to the statement that interacting individuals mutually and in virtue of the emergent interaction dynamics constitute their cognitive processes.[17] Social cognition as participatory sense-making then exhibits a *relational* kind of cognition. It is not to be located in either individual’s head, brain or even body, but *in between* interacting individuals. Note how this is fundamentally different not only from internalist, but also from externalist views of cognition and social cognition.

If this view is taken seriously, it has non-trivial theoretical as well as methodological consequences. Holding a proper metaphysical status (as an emergent macro-level property with causal effects on its parts, i.e., individual cognitive processes), interaction should be viewed as a proper *level of analysis*, since its dynamics are able to influence, constitute and enable individual mechanisms (cf. De Jaegher & Di Paolo, 2007, p. 491). Furthermore, it changes the *unit of analysis* for empirical as well as theoretical investigation. In order to find out about social cognition, how it is enabled and what its constitutive parts are, research would have to shift its focus from observing individuals to *individuals in an interaction* (thus including the interaction itself). Even the investigation of *either* individual of an interaction would still not be sufficient to explain social cognitive processes (or only a vanishingly low number of those), since the *in between* of individuals is thought to causally influence or even constitute the object of investigation.[18] Thus, the theoretical claim about the metaphysical status about interactions comes with methodological demands.

This also allows the assumption that interaction enables and even constitutes social cognitive processes *in virtue of* its metaphysical status. It is important to ask whether the notion of interaction as a constitutive element of a cognitive process would still work if interaction is viewed in a reductive (i.e., reducible to individuals) manner. In order to tackle this question, it should first be clarified why we would want to assign such a strong role to interaction to

[17] Note that in order to count as a social interaction, no coercion or exploitation of one individual by the other should be involved (De Jaegher & Di Paolo, 2007, p. 495).

[18] This is exactly what has been done by enactivists. The focus of their theories is on the description of interaction dynamics and the crucial elements such as timing, coordination, and synchronization.

begin with. Whether or not there really are cases of social cognition that are not explainable in individual terms is an open question, as will be discussed shortly.

2.3.3. The Interactive Brain Hypothesis

The view that interaction dynamics constitute behavior and cognition has been extended to the claim that they also play a fundamental role for brain mechanisms that are related to social cognition. This thought is captured by the ‘Interactive Brain Hypothesis’ (IBH), as described by Di Paolo and De Jaegher (2012). The hypothesis states that – whether there is an ongoing interaction present or not – the mechanisms that enable social understanding are derivative of functions of mechanisms that are used in interactions (cf. ibid., p. 2). In contrast to cognitivist views, the hypothesis claims that interaction does not merely yield inputs that are then processed and thus influence the outcome of cognitive processes, but that it actively and profoundly shapes the mechanisms themselves.

IBH comes in two complementary versions. The developmental interactive brain hypothesis (DIBH) states that

> [t]he functions of individual brain mechanisms involved in social understanding have been shaped during development by skillful engagements in social interactions where interactive processes have been involved in social performance in a more than contextual way.” (Di Paolo & De Jaegher, 2012, p. 5)

The contemporaneous interactive brain hypothesis (CIBH) claims that

> [e]ven in the absence of immediate interaction, the functions of brain mechanisms enabling social understanding are derived *contemporaneously* from functions used primarily in skillful social interactions where interactive processes are involved in social performance in a more than contextual way. (ibid.)

That being said, the authors emphasize that their claims do not mean that individual mechanisms play no role at all, but that “[t]he IBH simply makes the non-trivial proposal that among the necessary factors, we will always find some enabling or constitutive interactive elements.” (ibid.).

While the former claim (i.e., that interactive elements play an enabling role) rather uncontroversial, it remains unclear how the authors aim to argue for the latter. One interesting thought they elaborate on is a re-interpretation of the mirror neuron system. It could be assumed that interactive mechanisms serve as a function of its plasticity and thus shape social cognitive mechanisms. However, as is also mentioned by the authors, it is disputable whether observation could not yield the same learning effects (cf. De Jaegher & Di Paolo, 2012, p. 8).

In sum, according to IBH, the investigation of individual neural events does not yield a full explanation of what is going on in social situations. This is so because the cognitive work that needs to be done by either individual is not to be found internal to the skull, nor in anything external to the brain – but in the emergent relation between agents. This relation is the interaction pattern itself which is thought to carry an indispensable amount of the cognitive work load.

How are these claims justified? In what follows, I describe the perceptual crossing paradigm (PCP) which serves as the main empirical argument for the enactive stance.

2.3.4. The Perceptual Crossing Paradigm

While there have been numerous theoretical descriptions of social cognition in enactive terms, a solid empirical ground to back them up has often been claimed to be missing. In order to remedy this shortcoming, Auvray and colleagues (2009) developed the *perceptual crossing paradigm* I am going to scrutinize in this section. After describing the original set-up and motivation standing behind the experiment, I will summarize the interpretation of results.

Auvray and colleagues aim to investigate how perceptual activities between two individuals can be recognized 'directly'[19] and whether there are processes at play that can plausibly be ascribed to the interaction process itself. In the authors' words:

> The aim of the study reported here is to further investigate whether, in situations of perceptual interactions, some of the mechanisms underlying the recognition of others are intrinsic to the shared perceptual activity itself (i.e., intrinsic to the interdependence between the two perceptual activities). (Auvray et al., 2009, p. 34)

Behind that stands the motivation to explain the social ability to spontaneously ascribe mental states and thus to detect intentional beings in non-inferential, direct terms. The authors aim to show that perceptual encounters (e.g., catching someone's eye) can be framed

[19] "The subsequent question arises: does the recognition of a perceptual crossing necessarily require the concept of an intentional subject who possesses internal goals that are a pre-requisite for her action? Or, alternatively, would it be possible to consider the perceptual crossing as occurring more directly, i.e., prior to the elaboration of a complete theory of intentionality?" (Auvray, Lenay & Stewart, 2009, p. 34)

in enactive or interactionist terms. In a perceptual encounter, so it is claimed, interactions are created dynamically as an emergent property of the interaction itself. In situations of mutual perception, the two perceptual activities interact and thereby enable “the recognition of the intentionality of another person [which] is intrinsic to a shared perceptual activity.” (Auvray et al., 2009, p. 34)

The study was furthermore inspired by the double TV-experiment I described before. In the same manner, the still-face experiment (Adamson & Frick, 2003) shows that children get distressed and start crying almost immediately when their caregiver suddenly interrupts the interaction and shows no responses at all. In order to avoid the explanation that the infant’s cognitive machinery is already sophisticated enough to incorporate a complex cognitive process that involves several steps, the findings have been tried to be accounted for in interactionist terms that allow for ‘direct’ recognition of intentionality. The child seems to be, in any case, able to distinguish between situations in which the movements and actions of the other are or are not responsive to her own.[20]

Auvray et al’s perceptual crossing paradigm picks up the idea that there might be something inherent in the interaction dynamics that enables this recognition process. The set-up follows the principle of minimizing complexity for the sake of controllability. Pairs of two participants are seated in front of a screen they are unable to see (they are blindfolded). In one hand they hold a computer mouse which allows them to move the cursor (the ‘avatar’) along a line, in the other hand they hold a buzzer through which they receive a tactile feedback whenever their avatar encounters another object in the one-dimensional space. Participants can encounter three different objects, namely the other person’s avatar, a shadow of that avatar (the ‘mobile lure’) that follows the avatar at a fixed distance and a stationary object that remains fixed throughout the whole experiment. The task is to click the mouse whenever subjects think to have encountered the other’s avatar, thus to attribute the perceptual stimulus to the right kind of encounter:

> The task for participants was to recognize when the tactile stimuli they received were attributable to the encounter with the other participant’s avatar. Our principal hypothesis was that the interdependence of the two perceptual activities might be a sufficient factor to enable

[20] “The fact that children were able to distinguish a live interaction with their mother from a pre-recorded one suggests that the recognition of another person does not only consist of the simple recognition of a particular shape or pattern of movements, but also involves a property intrinsic to the shared perceptual activity: The perception of how the other’s movements are related to our own.” (Auvray & Rohde, 2012, p. 1)

> the participants to click more often after having met the partner's avatar (i.e., respond correctly) than after having met the mobile lure. (Auvray, Lenay & Stewart, 2009, p. 35)

It was found that in 65.9% of all clicks, the cause was a stimulation following an encounter with the other avatar, while 23% were attributable to encounters with the shadow and only 11% to those with the fixed object. However, when the percentage of all clicks was divided by the percentage of stimuli with a particular object, the results show that there is no higher probability for clicking after an avatar-avatar than an avatar-shadow encounter. This is due to the fact that the number of the first type of meeting is higher in total, thus amplifying the probability to click (cf. ibid., p. 39).

There are basically two findings that need to be accounted for, as the authors conclude from these results. The first is that participants were clearly able to distinguish between moving and fixed objects. The second finding is that participants seem to favor situations of avatar-avatar encounters (that is expressed in the higher number of that kind of meeting), while they seem unable to distinguish between the avatar and the shadow (which is obvious from the equal probability of clicking). As for the first, participants develop a particular strategy of moving the cursor, thus oscillating around the object they encounter. With this strategy, a fixed object can easily be identified, since the consecutive encounter follows within a reliable distance of time and is thus well predictable:

> This result leads us to propose a general a priori hypothesis: an object will be identified as being a moving object if the expectations appropriate to the perception of a fixed object are flouted, in one way or another. (ibid., p. 39)

It is further gathered that a moving object is far more likely to violate perceptual expectations, since it might elicit a stimulation although the participant did not move the cursor, or the stimulation could occur within an unexpected time span.

Following an enactive account of perception (Noë, 2004), Auvray and colleagues (cf. 2009, p. 44) suggest that the anticipation of sensory consequences was constituted by laws of sensorimotor contingencies. While it is acknowledged that this strategy is perfectly attributable to each individual, the second finding is said not to be explainable in individual terms and thus requires a non-reductive explanation at the level of collective dynamics. The reasoning goes as follows. It was found that participants reversed their direction of movement after encountering any object. Note that both subjects display the same oscillatory behavior around a source of stimulation, but that only when *both avatars meet,* both receive a tactile feedback. The result is that, according to the authors, "this co-dependence of the two perceptual activities thus forms a relatively stable dynamic configuration." (Auvray & Rohde, 2012, p. 3) The fact that an avatar-shadow encounter elicits feedback in only one subject is seen as not allowing for the emergence of a stable interaction pattern.

Couched in terms of DST, the authors refer to the interaction as eliciting a mutual attraction that is, due to its stability, the strongest attractor and thus explains the highest number of clicks:

> When the trajectories of the avatars cross, both participants receive a stimulation; if, as explained above, each participant then turns back, then they will meet again, and this pattern forms a relatively stable dynamic attractor. This co-dependence of the two perceptual activities contributed to favor the situations of mutual perception. (Auvray, Lenay & Stewart, 2009, p. 42)

Coming back to the study's goal of showing that the attribution of intentionality can be framed in interactionist terms, it is further claimed that both perceptual criteria and properties of the emergent interaction dynamics "favored the attribution of intentions to others" (ibid., p. 43). The authors take it that although participants had no intention of collaboration, the mutual perceptual activities gave rise to an attractor that then stabilized the interaction and thus explains the high number of avatar-avatar encounters.

It is exactly this point which motivated a follow-up study conducted by Froese and colleagues (2014). Picking up the claim by De Jaegher and colleagues (2010) that the paradigm proves that interaction can constitute social cognitive processes, they state that the original set-up by Auvray and colleagues as well as following modified versions (for a review, see Auvray & Rohde, 2012) have failed to prove this. This is so because in earlier experiments, participants were not explicitly asked to collaborate, thus missing the aspect of active and intentional co-regulation of their interaction (Froese, Iizuka & Ikegami, 2014, p. 2). Moreover, participants of the original paradigm should only be able to become aware of the other's presence when they succeeded in coordinating their interactions while the current task was changed so that participants can only succeed by mutually cooperating and coordinating their activities (ibid., p. 7).

The aim of the study was thus to evaluate the participants' ability to detect the presence of the other by actively and collaboratively engaging in an interaction. Subjects were paired up in couples and situated in the same set-up as in the original study, but this time with the explicit assignment to help each other to detect the other's avatar. The prediction was that "turning the individual epistemic task of detection into a social pragmatic task aimed at mutual coordination would implicitly facilitate discrimination of the other's avatar." (ibid.,

p. 3) The authors see this hypothesis confirmed in the *number of contacts* with common sensory overlap being the highest and that the *number of clicks* after avatar-avatar encounters was twice as high as for avatar-shadow encounters. It is concluded that:

> The near synchrony of recognition is due to the social interaction between the players; it cannot be explained in terms of independent entrainment to a shared time signal such as the duration of the trial […]. This indicates that social judgments were not so much based on an individual recognition of the other but rather on a mutually shared recognition of each other, i.e. on an interactively shared cognitive process. (Froese, Iizuka & Ikegami, 2014, p. 4)

Subjects were further asked to rate how clearly they felt the other’s presence. Again, the highest ratings occurred for situations of joint success (i.e., clicking after stimulation due to an avatar-avatar encounter).

This result is supposed to confirm the second prediction, namely that the engagement in an interaction leads to an increased feeling of presence and being in contact with the other. Thus, both hypotheses are deemed to be confirmed:

> Most participants of the current study were able to interactively coordinate their embodied interactions in the minimal virtual space so as to create sufficient conditions for jointly becoming aware of each other’s presence, and thus to click with higher accuracy. The fact that such co-regulation gave rise not only to a correct social judgment but also to an experience of the other’s presence supports the enactive approach to social cognition. (ibid., p. 8)

More generally, the results are taken to challenge cognitivist views of social cognition as a folk psychological capacity and furthermore to be plausible from an evolutionary and developmental perceptive “because an extendible mind can partially offload the mechanisms of cognition into its environment and thereby augment its capacities” (ibid.).

De Jaegher and colleagues (2010) take the original perceptual crossing paradigms as proof that interaction can indeed constitute social cognitive processes, since only the ability to discriminate moving from stationary objects can be attributed to individual mechanisms. The discrimination between shadow and avatar, however, can only be accounted for if interaction is taken to be a property that enables this task. The paradigm is thus probative of the potentially constitutive role for social cognition:

> The variation in the number of clicks is attributable only to the differences in the stability of the coupling and not to individual strategies. This experiment shows that the interaction process is not only enabling but plays a constitutive role. The phenomenon is a manifestation of the properties of the interaction pattern. (Froese, Iizuka & Ikegami, 2014, p. 445)

This statement has also implications for the validity of IBH. If it is true that in (some) social encounters, interaction patterns constitute part of the social cognitive process, then it indeed appears that social cognition cannot be accounted for by looking at individual brains. Instead, IBH claims that social cognition is an irreducible process which necessarily involves interaction patterns as part of it. This is supposed to be shown by PCP.

2.3.5. Evaluation

In order to see critically assess the just presented enactive position, let me start with the core assumption, namely that interaction forms a constitutive element of social cognition, leaving the phenomenon irreducible to individual processes.

One worry is that proponents of this conception confuse *enabling* with *constitutive* conditions. A first hint that this might be the case is found when looking at the taxonomy of possible roles of interaction for a social cognitive process X that De Jaegher and colleagues (2010, p. 443) have worked out:

> Accordingly, given X, and a particular situation in which X occurs: 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 is a constitutive element if P is part of the processes that produce X.

As Herschbach (2012) points out, however, it is rather unclear what exactly De Jaegher and colleagues judge to be a constitutive element. For additionally to the characterization given above, they also refer to it as a *part of the phenomenon itself*:

> A constitutive element is part of the phenomenon (it must be present in the same time frame as the phenomenon). The set of all the constitutive elements is the phenomenon itself. The *presence* of these elements is necessary, and therefore also enabling. (De Jaegher, Di Paolo & Gallagher, 2010, p. 443)

This ambiguity leaves us with two possibilities in which interaction can constitute social cognition: (1) it can either be among those processes that *produce* the phenomenon, but (if my interpretation of the quotation is correct) does not have to be a *part of the phenomenon*, or (2) interaction constitutes social cognition in the sense that it must be present at the same time as the phenomenon and is a *necessary part* of it.

Claim (1) describes a condition that should count as enabling, not constitutive. The idea seems to be that interaction *enables* a particular mechanism to *arise* in that it was present as a necessary part of the development of that skill, and therefore it should be called constitutive. This confuses the concepts profoundly and boils down to the assertion that interaction is an enabling condition and not that it constitutes a phenomenon in the sense that it is a part of it without which it would not exist.

Moreover, the view that being immersed in social interactions – especially from a developmental perspective – enables particular social cognitive skills can in principle be accounted for by any non-enactive theory that assigns a sufficiently strong role to extra-individual and situational contexts. Given that human newborns are completely helpless without a caregiver for an extraordinarily long time and given some rather anecdotal evidence of children which lacked interactive and emotional engagement in early development and had severe mental as well as bodily impairments (e.g., Zimmer, 1989), the fact that these contexts play a necessary role for social cognition seems almost trivial.

To shed more light on the second version of the claim (2), let me first point to a very helpful clarification of proponents of enactivist theories of social cognition. In their brief paper *Enactivism is not Interactionism,* De Jaegher and Di Paolo (2013) show that their statement is not that interaction dynamics *solely* constitute social cognitive processes without assigning any role to individual processes. Rather, their claim is that there are some instances of social cognition in which interaction is one of the constitutive factors – among others. In this sense, participatory sense-making describes a process which entails both individual as well as trans-individual elements, but the latter do not fully substitute the former (cf. ibid., 2013, p. 1).

With this clarification in mind, it becomes obvious that only if one accepts the depiction of interaction as forming an autonomous system (together with the individuals that are involved) in a strong enactive sense and thusly is seen as holding its proper ontological status as a system which actively steers the process of exploring the world, one can accept claim (2). This shows that in order to take an enactive perspective on social cognition, one needs to fully buy into their metaphysical assumptions about the status of relations between individuals, and individuals and their environments. I will discuss whether this is a desirable move in the following chapters.

The assumption that interaction forms a constitutive part of social cognitive processes also has consequences for the unit of analysis, as I argued above. Thus, IBH is proposed as an alternative to individualistic and internalist views prominent in (social) cognitive neuroscience. What is left to be explained by IBH, again, is what it actually means that interactions “can themselves enable socio-cognitive performance or even be a constitutive part of it” (Di Paolo & De Jaegher, 2012, p. 2).

At this point, remember the two possible interpretations of a constitutive element. The second, stronger one implies that interaction constitutes social cognition in the sense that it must be present at the same time as the phenomenon and is a necessary part of it. If this claim means that a particular neural mechanism is only recruited if the individual is in an interactive context, this view is perfectly describable if interaction is seen as an enabling condition. What would it mean, however, that interaction *constitutes* a neural mechanism? How can we plausibly conceive of a non-neural, relational process *being part of a neural* (i.e., internal) mechanism other than this process being among those that enable it?

One could plausibly say that some neural mechanisms only get recruited when the individual is in an interactive context. This claim, however, does not mean that interaction *constitutes* this mechanism, but that it is an important contextual component. The statement that “the neural mechanisms involved in social understanding acquire and sustain their current functionality thanks to past and present engagements in social interactions” (Di Paolo & De Jaegher, 2012, p. 2; whether in a developmental or contemporaneous sense) is a much weaker claim and would probably not be denied by opponents of the enactive theory of social cognition.

It appears that the main point of the authors boils down to the claim that interaction has a *more than contextual* role because neural activation is different depending on whether individuals are participating or have ever participated in an interaction or not. The crucial point then is to say that this is not so because interaction *as a context* yields different or more complex inputs, but because interaction dynamics have formed an autonomous system. As yet, there is no clear evidence of this hypothesis (cf. ibid., p. 7). The upshot of the paper seems to be that neuroscientific research has to pay attention to interaction dynamics because they can fundamentally alter both behavior and neural mechanisms.

There are several reasons the authors make this assumption, viz., that interaction can be an autonomous system, that there is a manifold spectrum of social dynamics like distribution and emergence of roles of participants in an empirical set-up (which marks the difference between truly *participating* in an interactive set-up or merely observing someone else), and the readiness of people to interact. The latter refers to the “disposition to engage or participate in socially meaningful situations” (ibid., p. 11) and is furthermore viewed in analogy to perception constituted by sensorimotor contingencies. Just as the perception of objects is said to yield affordances of potential ways to act upon the object in virtue of sensorimotor loops that constitute perception, it can be said that social situations afford possibilities of interaction, even if not actualized (Di Paolo & De Jaegher, 2012, p. 11). What is lacking to make their claims coherent, though, is a clear description of what would count as a constitutive element, and evidence which clearly shows that interaction indeed should be considered as such.

The perceptual crossing paradigm, aims to fill the empirical gap and is supposed to show that the results are best explained if it is assumed that the interaction relation between participants constitutively contributed to their behavior. There are basically three critical points in reply to that claim. First, proponents of this constitution-hypothesis have been accused of the coupling-constitution fallacy. Secondly, it has been claimed that the arising of collective dynamics can be explained in reductive individual terms, thus weakening the hypothesis. Finally, it is questionable whether the impact of the results of the paradigm can be said to be as great as has been claimed. Before I get to those points in more detail, let me say that, importantly, the first and second points focus on the strong claim that the paradigm reveals a *constitutive* role for interaction while they do not deny that collective dynamics play *some* role for the social cognitive processes in question.

To being with, scrutinizing these critical points requires a careful look at what the goals of the studies have been and what the interpretation really includes. There is, as has been pointed out by Auvray and Rohde (2012), a disagreement about the social cognitive process that is focused on by proponents and opponents of the constitution-hypothesis. Opponents claim that it must be the social *judgment* that the other's avatar has been encountered that is allegedly constituted by collective dynamics (Michael & Overgaard, 2012), since that is what was in the focus of the original paradigm. It is true that the *task* of the original paradigm indeed was to click when participants thought they met the other's avatar, thus including a social judgment of the presence of the other. However, the *outcome* that is accounted for by the constitution-hypothesis is something quite different, namely the high number of avatar-avatar encounters. What is said to be constituted by the collective dynamics is the "convergence toward the situations of perceptual interaction" (Auvray, Lenay & Stewart, 2009, p. 40) and not the judgment to be in one. De Jaegher and colleague's (cf. 2010, p. 445) interpretation of the results suggests that the process that is constituted by the interaction is the tendency to engage in social encounters. This is the phenomenon which is allegedly only explainable in reference to collective dynamics and which is "a manifestation of the properties of the interaction pattern" (De Jaegher, Di Paolo & Gallagher, 2010, p. 445).

After this clarification, let me now turn to the first critical point, namely the danger of committing the coupling-constitution fallacy. Michael (2011) and Overgaard (2012) state that in advocating the constitution-hypothesis, De Jaegher and colleagues are guilty of (or at least at high risk of) committing the coupling-constitution fallacy (Adams & Aizawa, 2008). The fallacy occurs when one element is seen as constituting the other, when in reality the two elements are merely coupled and stand in no constitutive relation to each other. Michael (cf. 2011, p. 567) criticizes that even though it could be said that the number of clicks (i.e., the outcome) is best accounted for by reference to the collective interaction dynamics, this does not mean that the interaction indeed constitutes the process that led to this distribution. He seconds the weaker claim that collective dynamics influence the outcome as a factor that is external to the individuals, but refrains from committing to the constitution-hypothesis. The reason is given in relation to the second critical point which claims that the results can be accounter for by non-enactive theories.

It has been claimed that the interaction dynamics that allegedly constitute the outcome can plausibly be said to be constituted by individual mechanisms:

> Our claim, in contrast, is that participants’ attempts to detect motion generate the collective dynamics which, in turn, enable participants to distinguish between moving and non-moving objects, and which also ensure that most encounters with moving objects will be encounters with the other participant’s avatar. (Michael & Overgaard, 2012, p. 298)

This follows from the strategy that is used by either participant and which allows for only one situation in which both participants’ avatars can be coupled, namely the very encounter of avatars. This fact does not, according to the authors, receive enough attention in interpreting the paradigm in enactive terms (Auvray & Rohde, 2012; Michael & Overgaard, 2012).

How can the results be explained without referring to ‘interaction dynamics’ in a strong enactive sense, viz., as an emergent macro-structure whose properties substitute a part of individual mechanisms? In order to answer this question, let us take a careful look at the paradigm. There are basically three situations either individual can be in during the task. Each situation differs with respect to the type of encounter. Within each encounter, as we have learned, individuals exhibit different behavioral patterns. Thus each situation is indeed different, but *in virtue of* the behavior of either individual. It could thus well be that individuals simply pick up subtle cues in the change of behavior of the other avatar, because the situation in which both participants receive a tactile feedback elicits a different kind of reaction. However, while this explains the *social judgment of the task*, it does not explain the *outcome of the task,* viz., the propensity to engage in avatar-avatar encounters. I propose that – if one does not want to accept the claim that interaction dynamics play that role – this outcome could be accounted for in virtue of the more general tendency of humans to engage in social situations.[21] If individuals are able to detect a social situation, they probably aim to engage in this one more often, simply because they favor them.

Finally, it is questionable whether the paradigm really succeeds to build a proper empirical basis on which grounds non-enactive theories of social cognition can be rejected as suggested by Auvray and colleagues (2009). The claim is that the paradigm reveals intention detection to occur without the need for inferential mechanisms, but it should be asked whether the experiment really tests for this cognitive process. It seems fair to say that it brings out the detection of biological motion. Intentions, however, are usually depicted as contentful mental states that can explain a certain behavior. Given that participants in the original study had to judge whether they encountered an object whose movements are guided by another person, the claim that this is due to intention detection seems misled. This relates to two last questions.

The first is whether the explanation of the results needs to be couched in enactive terms or whether they can be accounted for by a less radical theory. Froese and colleague’s (2014) claim that their results speak for an extendible mind that outsources parts of the cognitive work into the environment is, for example, compatible with an extended, yet non-enactive theory. The same holds for the claim that interaction dynamics influence the cognitive process. Especially the explanation of the ability to discriminate moving from fixed objects can well be expressed in a description of perception as picking up statistical regularities from the environment. Thus, if there is reluctance in the research field to fully commit to enactive views, it might stem from the fact that it is questionable what they genuinely add to already existing explanations.

The last question I want to address is in how far the paradigm can be seen as an empirical argument for turning towards enactive theories. This move, which will be described as the ‘interactive turn’ in chapter 4.1.2. in more detail, basically aims to bring extra-individual elements that play a role for social cognition into focus. It further demands that research

[21] For an extensive depiction of the importance and saliency of sociality in the light of evidence from SCN, see Lieberman, 2013.

should include a plethora of factors that are involved in a social encounter, such as emotional engagement, bodily processes, environmental and situational factors, individual mechanisms, the relation between interlocutors, and so on. This of course involves a challenge for experimental set-ups to better reflect ‘real-life’ situations, thus allowing participants to move freely, engage with each other and not feel like in an artificial environment that could not be found in real life.[22] However, the experimental set-up of the paradigm is neither close to a real-life situation, nor does it involve much movement (the only movement that is required is that of moving a computer mouse and clicking its left button) or emotional engagement. Also, sensory stimulation is rather low and does not reflect the high complexity of a social encounter. It is clear that – at least with the methods currently available – higher controllability comes with lower ecological validity (and *vice versa*). Still, the perceptual crossing paradigm seems rather far away from the latter.

While the paradigm can be said to elucidate the importance of interactive contexts for the social process of motion (and, probably, agency) detection, one should be cautious to overestimate its scope. It should also be considered that the emphasis on interaction, external contexts, emotions and bodily processes, and the importance of these factors for social cognition is nothing that proponents of non-enactive theories are committed to reject.

After this extensive criticism, let me point to important contributions of the enactive view to the debate on social cognition, more of which will be discussed in chapter 4.1.3. Although the phenomenological perspective already raised awareness about the fact that mindreading approaches have neglected social interaction, it is an achievement of proponents of enactivism to stress the importance of interactive situations. Even though researchers have been reluctant to adopt the radical metaphysical claims of interaction as a constitutive element, there is an increasing effort to implement interaction scenarios into empirical designs (Schilbach et al., 2013). Given that interactive situations can at the least be seen as enabling and contextually embedding some social cognitive mechanisms, I shall thus pick it up as another component of social cognition:

[22] Obviously driven by the motivation to account for this, the abstract of the original perceptual crossing paradigm starts with the following question: “How in real life or through the use of technical devices can we recognize the presence of other persons […]?” (Auvray, Lenay & Stewart, 2009, p. 32)

c_8 – interaction
Interaction is seen as an enabling and contextual component of social cognitive processing.

The enactive perspective has indeed elicited much debate about the status about interaction, but also about the role of the body for social phenomena. Where this debate is currently leading will be scrutinized in chapter 4.

3. Social Cognitive Neuroscience: Empirical Findings and Conceptual Challenges

> *Although social neuroscience needs to be broad, it also needs a focus for nucleation, otherwise it threatens simply to merge with cognitive neuroscience or splinter into an array of otherwise unrelated projects. And of course, there is a focus: it is the word ''social'' that is raising questions about how best to circumscribe this term. (Stanley & Adolphs, 2013, p. 817)*

After having focused on philosophical and phenomenological approaches to social cognition, I now turn to the more empirical domain of the wide research field. In this chapter, the goal is to scrutinize and discuss the aims, targets, and challenges within social (cognitive) neuroscience (SN/SCN). While philosophy and phenomenology started to investigate social phenomena rather early, SN/SCN did not get off as a proper research field before the 1980s. To describe reasons for this late development and to show that once it started, SN/SCN quickly moved on to form a genuine and genuinely interdisciplinary domain of research is one main goal that will be pursued in chapter 3.1. This is an important aspect since it motivates the view that the field needs a proper theoretical background to couch its findings. In chapter 3.2., one important conception of SN/SCN is looked at in more detail, namely the so-called 'social brain'. After clarifying this concept, I proceed to depict four core networks that are supposed to underlie social cognitive processing and show how their investigation influenced theoretical considerations.

Lastly, in chapter 3.3., I focus on two conceptual challenges in the field. The first concerns the issue of whether there is something special about social cognition which differentiates it from general cognition. I present several attempts to tackle this question and name consequences of possible answers. Lastly, I ask what kind of terminology the interdisciplinary nature of SCN requires and argue that 'terminological consistency' still holds as one desideratum for a theory of social cognition.

3.1. The Historical Growth of a Genuine Research Field

Social (cognitive) neuroscience (SN/SCN) is a relatively new research field which gathers researchers from a variety of disciplines to find out about social cognition. There are several reasons to assume that it indeed makes sense to view SN/SCN as a genuine research field.

This is especially interesting when considering its interdisciplinary origins. The goal of this chapter will be to argue in favor of the depiction of SN/SCN as a *genuine* and *genuinely interdisciplinary* research field:

(1) Combining already existing methods and concepts from cognitive psychology, social psychology and neuroscience with a novel acknowledgement of social cognition as a proper domain of research, the field of SN starts to grow. In chapter 3.1.1., I present the principles and aims underlying SN.

(2) Although inherently related to SN, the field of social *cognitive* neuroscience arose when the focus shifted from the biological to the cognitive level of description. Chapter 3.1.2. discusses goals, targets and theoretical considerations of SCN.

3.1.1. The Rise of Social Neuroscience

Of course there have been investigations in the field of neuroscience and social psychology that would fall under the category of social neuroscience before it became an acknowledged and highly organized field of research. However, as mentioned in many reviews on the topic, it was not before the late 1980s that social neuroscience started to formulate explicit goals, hypotheses and grew to become a distinct discipline (Lieberman, 2012; Ochsner, 2007). The rise of the field came with an increased interest in socioemotional processes in humans. As Ochsner (2007) shows, social as well as emotional components of behavior were out of the scope of neuroscience for several reasons. Not only was the methodology at hand not developed enough for thoroughly studying humans – before imaging technologies such as functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) allowed studying healthy humans, researchers had to rely on animal models, postmortem examination or cases of lesions or diseases. Also, socioemotional behavior was considered as rather animalistic, low-level abilities that contradicted the image of humans as rational beings. However, with Le Doux, Panksepp and Damasio (amongst others), work on emotion and social behavior became more popular.

Ochsner (2007) furthermore describes the development of related research fields, such as cognitive neuroscience in the late 1980s, investigating complex behavioral phenomena with neuroscientific methods, affective neuroscience and, finally, social neuroscience. The growing interest in social behavior also manifested itself at a conceptual level: "For example, the term "social cognition" came into common usage in the early 1980s to refer to the use of cognitive psychological theories and methods to study phenomena typically of interest to social psychologists." (Ochsner, 2007, p. 42) Thus, in the research field of social neuroscience, the scopes of different, already existing disciplines – neuroscience, social psychology and cognitive psychology – merged to reach the goal of social neuroscience, which has been described as follows:

> Social neuroscience emerged in the early 1990s as a new interdisciplinary academic field devoted to understanding how biological systems implement social processes and behavior, capitalizing on biological concepts and methods to inform and refine theories of social processes and behavior, and using social and behavioral concepts and data to inform and refine theories of neural organization and function. (Cacioppo & Decety, 2011, p. 5)

It is assumed that social cognition needs particular neural, hormonal, cellular and genetic support to evolve and thus social neuroscience is furthermore described as "the study of the associations and influences between social and biological levels of organization." (ibid., p. 3)

The term 'social neuroscience' was first used in a seminal paper by John Cacioppo and Gary Bernston (1992) with the title *Social psychological contributions to the decade of the brain: Doctrine of multi-level analysis.* The authors claim that the increasing attention to the brain (e.g., the official declaration of "the decade of the brain", ibid.,p. 1020) calls for an increasing awareness of the fact that brains are embedded in an environmental context, which – for most species – is a highly social context. Therefore, social psychology as a science that focuses on complex social behavior needs to be seen as a fundamental and necessary complement to neuroscience. The importance of social psychology for neuroscience furthermore becomes apparent when considering the fact that neurochemical events influence social behavior and *vice versa*. Cacioppo and Bernston (1992) hence argue that social neuroscience needs a multi-level, non-reductive approach in order to do justice to the complex nature of social behavior and its underlying mechanisms.

The question which level of analysis is appropriate to describe social cognition at remains open until today. Importantly, the scientific community of social neuroscientists seems to agree on the goal of social neuroscience to *find bridging principles* between the different levels of analysis (e.g., Kennedy & Adolphs, 2012; Matusall, Kaufmann & Christen, 2011; Ward, 2012). The issue of levels of analysis also implies the question of reductionism. How reductive should an account or a theory of sociality be? The question is important for research and I will now describe Cacioppo's and Bernston's (1992) proposal, since it yields some interesting thoughts on how to deal with these problems.

Basically, the authors argue for two claims. Firstly, although the brain and biological mechanisms are essential for social processes, a *comprehensive* account of social behavior cannot be yielded in a reductive manner.[1] This does not imply – secondly – that there cannot be plausible reductive explanations of lower-level phenomena in principle. The crucial point is that *complex* behavioral phenomena need a *complex* analysis at different levels that then is merged into a *comprehensive* account by stating bridging principles.[2] Through "multilevel integrative analysis" (Cacioppo & Bernston, 1992, p. 1021), such an account can be developed.

The authors proceed by describing a spectrum of levels of organization that reaches from microscopic, lower levels of organization (e.g., cells, molecules) to macroscopic, high levels of organization (e.g., sociocultural aspects). While neuroscience can yield reductive explanations at the lower levels, social psychology aims to investigate the higher end of the spectrum. Since there is no isomorphism between the conceptual and descriptive units of neuroscientific and social psychological investigation, it is the task of *social* neuroscience to scrutinize relational aspects of multiple levels. This is what the authors call a multilevel integrative analysis:

> Thus, by multilevel analysis we mean the study of a phenomenon from various structural scales or perspectives, ranging from the neuroscientific ("microscopic") to the social psychological ("macroscopic"). By integrative, we mean simply that analyses of a phenomenon at one level of organization can inform, refine, or constrain inferences based on observations at another level of analysis, and, therefore, can foster comprehensive accounts and general theories of complex psychological phenomena. (ibid.)

The underlying "doctrine of multilevel analysis" (ibid., p. 1023), is characterized by several principles. These principles, which I will shortly present in the following, have been picked

[1] Please note the distinction between reductionism and substitutionism that Cacioppo and Bernston (2004, p. 108) emphasize: "*Reductionism* embraces the ability to relate one level of organization (e.g., the social) to another (e.g., the hormonal), but recognizes that causal links between levels go both ways, and that lower levels of analysis can never entirely replace or substitute for higher-level analyses. The opposing construct of *substitutionism* holds that one level of analysis (generally a higher) can be replaced or supplanted by another level (generally a lower), and the goal of science is the pursuit of explanations at the lowest possible level of analysis."

[2] "In summary, reductionism has contributed to the solution of the most perplexing scientific problems in human history […] and has much to contribute to our understanding of social and psychological phenomena. However, it is counterproductive to presume that reductionism will convert the abstractions of the psychological sciences to "real" science in the coming millennium, just as it is counterproductive to presume that reductionism produces insights that are irrelevant to theories of social processes and phenomena." (Cacioppo & Berntson, 1992, p. 1026)

up in subsequent descriptions not only of social neuroscience, but also social *cognitive* neuroscience (see chapter 3.1.2.).

The first is the 'principle of multiple determinism', stating that one event token at one level of analysis can be influenced and determined by several precedent events at different other levels of organization. The 'corollary of proximity' complements the first principle and claims that "the mapping between elements across levels of organization becomes more complex (e.g., many-to-many) as the number of intervening levels of organization increases." (Cacioppo & Bernston, 1992, p. 1023) The second principle, the 'principle of nonadditive determinism', deals with the relation between wholes and their parts, asserting that in order to determine properties of collective wholes, properties of their parts need to be studied thoroughly across levels. Lastly, the third 'principle of reciprocal determinism' (ibid.), describes that there can be "mutual influences between microscopic (e.g., biological) and macroscopic (e.g., social) factors in determining brain and behavioral processes." (ibid.) From these assumptions it follows that a purely reductionist account at a neurophysiological level can mask important insights for a comprehensive theory of social behavioral phenomena and that interdisciplinarity is *necessary* to avoid this shortcoming.

The most important insight here is, it seems to me, that the research field on social cognition acknowledged its manifoldness at a very early stage. As we will see repeatedly throughout this thesis, this insight will be the key for answering the overarching question of this thesis, viz., what kind of theory is needed to appropriately describe social cognition? The quest for a multi-level analysis has already been mentioned as one central desideratum for a theory of social cognition. This is also, as will become more obvious in later chapters, one constant source of challenges and conflict. The question of how different levels of analysis relate is pressing and finding tentative answers will be of major importance; mainly because it will facilitate the dialogue between disciplines.

3.1.2. The Emergence of Social Cognitive Neuroscience

In 2001, Kevin Ochsner and Matthew Lieberman announced *The Emergence of Social Cognitive Neuroscience* in their same-titled paper. Together with a growing number of publications, conferences and journals that were dedicated to this newly emerging field, ever more researchers started to identify themselves as social cognitive neuroscientists (Singer, Wolpert & Frith, 2004). They came from different disciplines, including social psychology, developmental psychology, cognitive neuroscience, affective neuroscience and social neuroscience (Ochsner, 2007). The term 'social cognitive neuroscience' was used to refer to the study of socio-emotional phenomena, combining questions and methods from (mainly) social psychology and cognitive neuroscience, which Ochsner calls the "parent disciplines" (ibid., p. 39) of SCN. In this section, I wish to show that there are several aspects that speak in favor of viewing the field as a *genuine* and *proper* research field.

Although there already was a research field (SN) whose goal it was to link social behavior to neurobiological processes, the rise of SCN found many followers who did not fully identify with either of SN's goals and aimed to focus on social *cognitive* phenomena rather than being interested in the more general relation between social behavior and biological phenomena. Ochsner (ibid.) reviews the historical process by taking the development of SCN's 'parent disciplines' into account. According to him, social psychology underwent a change in shifting its focus from the social level down to the cognitive level. Cognitive neuroscience was born when cognitive psychology moved down from focalizing information processing to neural processes. Here is the author's description of how this process ultimately led to the formation of SCN:

> By taking a step down, social psychology became social cognition, and cognitive psychology became cognitive neuroscience. Thereafter, each field took many conceptual and empirical steps forward. SCN's emergence can be construed as either other a step down for social cognition or a step up for cognitive neuroscience. (ibid., p. 44)

While social psychology traditionally focused on broad topics like the alternation of perception and behavior in the presence of other people, research in cognitive neuroscience investigated neural systems that were thought to underlie and enable certain behavioral output. Their mutual interest lies in examining cognitive phenomena and their relation to social behavior.

As collaborations of researchers increased, SCN was born (Ochsner, 2007). Reviewers of the process emphasize that this amalgamation required both sides to get to know and consider the principles of the other's perspective to avoid restricted or naïve theoretical and methodological inquiries (Ochsner & Lieberman, 2001). Ochsner (2007, p. 43) summarizes the interdisciplinary merger:

> Thus, the term "SCN" appealed to researchers who (1) were interested in using cognitive neuroscience methods to study a wide array of socio-emotional phenomena, (2) wanted to use this combined methodology to elucidate the information processing level of analysis, and (3) did not identify with the types of research questions and content areas previously associated with related fields, such as SN [social neuroscience], AN [affective neuroscience], and CN [cognitive neuroscience].

The quotation furthermore highlights several aspects worth mentioning. First, by dissociating themselves from related disciplines and the will to define their own scope, researchers of SCN take an important step towards creating a *genuine* research area. Although SCN certainly got off as a *combination* of already existing methods, theories and questions, it merged these aspects in a *unique* way so that novel and fresh scopes arose.

Second, the definition of levels of analysis was one important step in defining SCN's perspective. While cognitive neuroscience adopted Marr's model,[3] and social psychology's focus lay on the socio-behavioral level, SCN combined these to a three-layered model of levels of analysis: The *social* level includes descriptions of a person's experience of social situations and her behavior in social settings.[4] At the *cognitive* level, psychological processes that lead to a certain phenomenon are scrutinized in terms of information processing, while the *neural* level investigates brain mechanisms that are thought to underlie those social cognitive phenomena (Ochsner & Lieberman, 2001; Singer, Wolpert & Frith, 2004).

Third, the focus of SCN has been described as resting on the cognitive level, since this is where the largest overlap of interest and expertise of its parent disciplines lies. SCN is thusly characterized by Ochsner and Lieberman (2001, p. 719):

> The name social cognitive neuroscience denotes both the interdisciplinary nature of the field and its emphasis on integrating data from multiple levels of analysis, ranging from experience and behavior of motivated individuals in personally relevant contexts (the social level) to these information-processing mechanism that give rise to these phenomena (the cognitive level) to the brain systems that instantiate these processes (the neural level).

Similar to SN, SCN aims to provide descriptions at multiple levels and link them. The previously mentioned desideratum 'multi-level analysis' can thus be found again. Further, it seems that the emergence of the research field rests upon the assumption that there is something distinct about social cognition which justifies its formation. Instead of treating social cognition as just one case of cognitive processing, SCN was motivated by the conviction that there is something so special that is deserves its own, genuine research field.

[3] Marr (1982) claims that there are three levels of description, namely computation, algorithms, and implementation. As Ochsner (2007) describes, Marr not only believes these levels to function independently, but also frames his views in a functionalist perspective. Thus, any computation can be produced by any algorithm and in principle be implemented in any hardware. At this point, CN disagrees with Marr, since its target phenomenon is not any possible hardware and how it implements computation in multiple ways, but more specifically the *human brain*.

[4] Importantly, it has been claimed that this also includes behaving or interacting with oneself, hinting to the still-prevalent interest in the self and its relation to others (Ochsner, 2007; Singer, 2012).

Further, the goals and attempts of the research field of SCN are defined as follows:

> SCN research uncovers relationships between variables described at these three levels of analysis by conducting studies that provide information about the psychological processes associated with specific brain systems, or uses information about brain systems to inform theories of the psychological processes engaged in social behavior. (Ochsner, 2007, p. 51)

Thus, the ultimate aim of SCN is not the search for social modules in the brain or mere brain mapping. Rather, the field wishes to provide more than correlations of brain processes and behavioral output by using the findings to improve theoretical progress. However, as is emphasized by many researchers in the field, “that can only happen when researchers in the field have built a baseline of knowledge about the brain systems underlying specific types of social or emotional processing.” (Ochsner & Lieberman, 2001, p. 725) To find these brain systems has been one central goal in SCN, and I will thus now turn to the heart of the field, namely the ‘social brain’,

3.2. The ‘Social Brain’

One basic intuition in both social neuroscience and social cognitive neuroscience is that there is such a thing as the *social brain*. The term captures several ideas. In general, it expresses that there are regions in the brain that subserve social cognitive processing. It also entails claims about the role of social cognition for the evolution of the brain and, more specifically, makes assumptions about how the architecture of brains relate to social cognitive abilities. In this chapter, I review findings on the so-called social brain and in doing so proceed as follows:

(1) In chapter 3.2.1., I introduce the Social Brain Hypothesis (SBH) and describe several versions of it. It will turn out that although the concept of a social brain was used to refer to the evolutionary development of the brain for being social, it is used nowadays to refer to those regions of the brain which underlie social cognitive processing.

(2) This modern view is scrutinized in chapter 3.2.2., where I discuss the concept of a social brain briefly and give reasons for my further proceedings of the chapter.

(3) In chapters 3.2.3.-3.2.6., four core networks of the social brain are described, namely the social perception network (3.2.3.), the mentalizing network (3.2.4.), the mirror neuron network (3.2.5.), and the empathy network (3.2.6.). Here, I focus on their

alleged functions and show how research on these networks has been exploited theoretically.

3.2.1. The Social Brain Hypothesis (SBH)

The term 'the social brain' has been coined by Michael Gazzaniga (1985). In his book *The social brain: discovering the networks of the mind*, the author aims to show how lateralized functions of the right and left hemisphere contribute to social processes. Although lateralization is no main topic in social neuroscience anymore, the book can be seen as the "first modern attempt to explain the emergence of social psychological phenomena in terms of the organization or the function of the brain." (Lieberman, 2012, p. 432)

In 1990, Leslie Brothers (1990) picked up the term and hypothesized that there was a confined set of regions in the brain that was specialized for the processing of social stimuli. Robert Dunbar (1992) found that there is a correlation between the size of the neocortex in relation to the whole brain and the group size a species lives in. This finding is interpreted in favor of the idea that the demands of dealing with a large social group require a big brain. Thus, humans and other primates that live in relatively large social environments needed to evolve a larger brain in order to be able to deal with the correlated challenges.

Another version of the so-called *social brain hypothesis* (SBH; Dunbar, 1998) makes further assumptions about the evolutionary role of social cognition as a deceiving mechanism. The *Machiavellian Intelligence Hypothesis* (Whiten & Byrne, 1997) states that "[…] the competition for social skills led to the evolution of cognitive mechanisms for outsmarting others, and fuelled the expansion of the human brain and perhaps the elaboration of certain neural systems." (Adolphs, 2003c, p. 166)

While these views try to explain the *size of the whole brain* in relation to demands of a complex social environment, SBH can also be seen as the idea that there are regions in the brain that are *specialized* for social processing. The term 'social brain' then describes sets of regions that are confined to process social stimuli or otherwise enable individuals to deal with their conspecifics. It is this conception that (most) researchers nowadays refer to when talking about the 'social brain'.

Note how this notion of the social brain relates to modular views of the mind and the brain. Gazzaniga (1985), for example, heavily relies on Fodor's theory of modularity. Indeed, at first sight it seems plausible to adopt a theory that emphasizes domain-specificity in order to argue for the claim that there are networks that evolved for a certain type of cognition.

Furthermore, as already mentioned, impairments of social cognitive abilities – for example the case of Phineas Gage (Damasio, 2006), ASD (Frith, Morton & Leslie, 1991), Williams Syndrome (Järvinen-Pasley et al., 2010), amygdala lesions (Adolphs, 2003b) – are rather circumscribed, leaving most other cognitive functions unscathed. This could speak for both encapsulation and neural localizability, features that are predicted by a modular view of the mind.

The idea that sociality is to be found 'in the brain' gained more and more attention in the 1990s as scientists started to discover ever more regions that strongly correlated with the presentation of social stimuli. In the following, I will review the findings of the core regions for social cognition.

3.2.2. Networks for Social Cognition

The goal of the following sections is by no means to comprehensively review the plethora of work that has been done on the social brain in SCN. Rather, I want to give a very rough sketch of the most important neurobiological facts and then focus on aspects that are of greater philosophical interest.

The term 'social brain' is used throughout the literature in SCN to refer to a set of regions and networks in the human brain that have been hypothesized to underlie social cognition. Research in the past two decades has revealed many of such regions and networks and has shown in numerous replications and follow-up studies that these quite reliably correlate with the processing of social stimuli.

There are many possible ways to categorize and describe these areas. Here I want to pick up the description of four core networks: the social perception or amygdala network, the empathy network, the mentalizing network and the mirror neuron network (Kennedy & Adolphs, 2012; Stanley & Adolphs, 2013). This view offers the possibility to include quite a wide array of components of social cognition, but also allows for restrictions. After giving a short description of empirical aspects of the network under scrutiny, I will focus on a set of questions that are of interest in this thesis. They concern the conceptual and theoretical framing of empirical results, but also whether these findings were used to derive new theoretical ideas or to revise existing theories. I also focus on how the mapping of neural and psychological processes is depicted and whether there are underlying background assumptions that are of importance.

In depicting the ‘social brain’ in terms of networks, a rather recent development in SCN is shown. While research in the early phase of the field focused on single brain structures and how they might implement social functions, the scope has widened and shifted to consider sets of structures as networks for social cognition (Frith, 2007; Stanley & Adolphs, 2013). This might well be due to the growing rejection of modular or phrenological views of the brain since the advent of alternative models of the brain that advocate a more dynamical perspective. In fact, Forbes and Grafman (2013) describe it as one major future direction for SCN to integrate more dynamic views of the brain to be able to consider a wider array of external and internal influences.

3.2.3. The Social Perception Network

At the center of the social perception network is the amygdala and it is hence also known as the amygdala network (cf. Stanley & Adolphs, 2013, p. 821). In general, the amygdala can be seen as a structure that functions to evaluate emotional components of social stimuli, such as facial expressions, and thus contributes to judging social situations. More specifically, the structure seems to process rather negative emotions, such as fear in threatening or dangerous situations. Adolphs (2001, p. 233), whose research focuses on the role of the amygdala for social cognition, hypothesizes that its primary function is the “linking of perceptual representations to cognition and behavior of the emotional value of the stimuli.” In several studies that tested patients with amygdala lesions against healthy subjects this hypothesis was tested and confirmed.[5]

One famous study, using fMRI and CT (computer tomography), showed that S.M., a woman with bilateral amygdala lesion, judged negative emotional expressions such as fear or anger drastically different than control subjects (Adolphs et al., 1994). The subjects were presented with pictures of people who display different emotions and then were asked to rate the emotional value. In a similar study, the same group of researchers found that patients with amygdala lesions judged faces of unfamiliar persons as more approachable and trustworthy as did healthy controls (Adolphs, Tranel & Damasio, 1998).

Another hint to the role of the amygdala for processing subtle social cues comes from a lesion study in which patients were presented with the Heider-Simmel paradigm (Heider &

[5] For a review, see Adolphs, 2003b.

Simmel, 1944; Heberlein et al., 2000). In their famous experiment, Heider and Simmel showed subjects a video of geometrical figures that move in certain ways. When subjects were asked to describe what they saw, they all put their descriptions in terms that are also used to describe human behavior. Although the figures that are seen do not show any similarity to humans – they are simply triangles, squares and circles in different sizes that move in relation to each other – their movements seem to be sufficient to elicit a quite human-like description. Here is one example:

> Triangle number-one shuts his door (or should we say line) and the two innocent young things walk in. Lovers in the two-dimensional world, no doubt; little triangle number-two and sweet circle. Triangle-one (here-after known as the villain) spies the young love. Ah! ... He opens his door, walks out to see our hero and his sweet. But our hero does not like the interruption (we regret that our actual knowledge of what went on at this particular moment is slightly hazy, I believe we didn't get the exact conversation), he attacks triangle-one rather vigorously (maybe the big bully said some bad word). (Heider & Simmel, 1944, p. 247)

While controls thus described the abstract visual stimuli in merely social terms, S.M. had no problem putting what she saw in perfect geographical terms:

> OK, so, a rectangle, two triangles, and a small circle. Let's see, the triangle and the circle went inside the rectangle, and then the other triangle went in, and then the triangle and the circle went out and took off, left one triangle there. And then the two parts of the rectangle made like an upside-down V, and that was it. (Heberlein et al., 2000)

The quotation shows that S.M. does not, like controls, attribute intentional behavior to the visual stimuli automatically. Thus, the amygdala seems to be crucial for the processing of biological motion in social terms.

Furthermore, the structure is highly connected to other regions that are important for perception of biological motion and faces, superior temporal sulcus (STS) and the fusiform gyrus. While the fusiform gyrus processes more static features of faces, STS is for changeable features as well as biological motion (Allison, Puce & McCarthy, 2000). Both structures are found in the temporal lobe, which provokes the question whether face perception is processed in a domain specific way in this region. Adolphs (2001) argues that the question whether social cognition draws upon domain specific mechanisms or not could be answered both ways. It might be that the fusiform gyrus and STS evolved for processing faces, but are also active when computationally similar phenomena are presented.

I will now focus on a re-interpretation of the role of the amygdala in light of new empirical evidence. The structure was thought to link social stimuli with representations of their emotional value and saliency, as described above. A set of lesion studies with patient S.M.

has led to a revision of this view. Interestingly, the new interpretation of amygdala's role has been tried to be embedded in externalist views of cognition, or at least has been used to emphasize that social cognition most probably involves an external component (Adolphs, 2006a). Since this attempt not only reflects the growing consideration of externalist views of (social) cognition, but also nicely exemplifies how the main cognitive load is still thought to be internal, I will concentrate on the paper *How do we know the minds of others? Domain-specificity, simulation, and enactive social cognition* (ibid.).

The basic claim in this piece of work is "that social cognition involves loops of processing that are extra-neural. It involves the bodies, and the social environment, in which brains are embedded." (ibid., p. 27) The first step in arguing for this claim consists of criticizing fully internalist views of social cognition for overestimating the role of the brain and underestimating the role of external components. The alternative picture that Adolphs draws depicts social cognition as consisting of three mechanisms, one for perceiving social stimuli, one for inferring hidden information and finally "mechanisms for exploring the social environment and probing it interactively." (ibid.)[6] This means that the three processing stages of social cognition – perception, cognition, behavior – do not necessarily work in that order. Not only does cognition lead to behavioral output, but also can behavior be seen as a means to actively gather new information. Note how this stirs up the sandwich view of the mind and at least potentially gives up the notion of perception and action to be peripheral to cognition.

To exemplify this point, the new interpretation of the role of the amygdala is contrasted with the old version. In that version, the degree of embodiment was reflected in that the amygdala was thought to link *somatic* with visual representations. Indeed, this is one famous view of 'embodiment' in cognitive science more generally, namely that cognition is embodied in that it draws on representations of bodily information. However, the new interpretation allots the amygdala a more active role. Let me briefly summarize how researchers came to this conclusion. As a follow-up on earlier studies that hinted to impaired fear recognition in patient S.M. due to her bilateral amygdala damage, the patient was presented with fearful faces in a set-up that was supposed to give more hints to the underlying mechanisms. Using eye-tracking devices, it was revealed that S.M. did not – as control subjects – focus on the

[6] The author suggests that the degree to which these three processing components are integrated could be one defining feature to make social cognition special (Adolphs, 2006b).

eye-region of the face in the photo she was presented with, but rather focused on nose and mouth (Adolphs et al., 2005). However, when she was *explicitly instructed* to fixate on the eye-region, her impairment of recognizing fear in faces vanished. This reversion, it is claimed, shows that S.M. is indeed able to use information of the eye-region and thus to recognize fear. What is impaired is her ability to fixate on the area in the face where this information is available. Drawing on these findings, it is proposed that the amygdala "modulates other brain structures to enhance the processing of stimuli about which more information needs to be acquired." (Adolphs, 2010b, p. 48)

While this interpretation sounds rather internal, Adolphs (2006a, p. 27) offers another description: "The subject, as a result of damage to the amygdala, lacked a normal mechanism to explore the environment." Although this is called an "enactive aspect" (ibid.) of social cognition, it reflects a rather moderate version of embedded and embodied cognition. The main cognitive load is still considered to be carried by neural structures, a view that is strengthened by reference to lesion studies that show the causal role of the brain for cognitive processing.

To avoid undermining the role of the body and environment, it is suggested that social cognition could involve both internal modeling and external probing. This is – as mentioned by Adolphs – closely related to accounts of situated cognition. These accounts are described as claiming that instead of having ready-made representations of the environment, human agents have the ability to actively seek information in the outside world. Note that these considerations are all fully compatible with internalist, functionalist and representationlist views of the mind and as such do *not* qualify as enactivist (see chapter 2.3.). That the claims made in the paper under scrutiny do not require a strong externalist background theory is also reflected in the terminology that is used. Mentalizing is still thought to be the central capacity that is necessary to infer the other's hidden mental states. These are defining features of theories that are rejected by enactivist and interactionist views of social cognition. Still, by taking seriously the importance of actively seeking salient stimuli and the possibility of extra-neural trajectories in doing so, Adolphs opens the way to further theoretical investigation that takes into account components of social cognition that entail embodiment (c_7), and probably even interaction (c_8).

3.2.4. The Mentalizing Network

The term ‘mentalizing’ (as described in more detail in chapter 1.2.6.) refers to the ability to infer the mental states of other people. Oftentimes, mentalizing and ToM are used interchangeably in the majority of the literature. In general, the following brain structures are correlated with mentalizing: STS, TPJ (temporo-parietal junction), PC (precuneus) and MPFC (medial prefrontal cortex) (Frith, 2007; Frith & Frith, 2006, 2012; Koster-Hale & Saxe, 2013; Mitchell et al., 2006).

However, the exact roles of these regions for mentalizing remain controversial. While most studies suggest a role for MPFC in mentalizing, there is also one report of a patient with lesions in MPFC that does not show crucial impairments in these tasks (Bird, 2004). Saxe (2006a) argues that there is no clear evidence for MPFC to play a necessary role in reasoning about contents of mental states. She is also rather cautious with drawing conclusions on the decisive role of other structures thought to underlie ToM. While Saxe claims that the activation of TPJ as shown in fMRI studies does rely on reasoning about other people’s minds, she emphasizes that the exact role of the area is still unclear. The author also warns to assume too much homogeneity for both ToM as a capacity and the underlying regions as functional units (Saxe, 2006b).

It is somewhat unclear, furthermore, whether or not mentalizing is a conscious, meta-cognitive ability or also refers to more unconscious inference mechanisms. In this subchapter, I will thus describe Frith and Frith’s (2012) view which divides mechanisms of social cognition into implicit and explicit processes within a dual-process model. Furthermore, the implicit/explicit distinction is translated into a distinction between cognitive and *meta-cognitive* mechanisms (Frith, 2012; Frith & Frith, 2012).

In an attempt to clarify these concepts, meta-cognition is defined as the “reflection of mental states, including own mental states (introspection); others’ mental states (popular psychologizing); mental states in general (philosophy of mind)” (Frith & Frith, 2012, p. 289). Mentalizing is described as the “implicit or explicit attribution of mental states to others and self (desires, beliefs) in order to explain and predict what they will do.” (ibid.) Explicit mentalizing then is a meta-cognitive ability that can also be referred to as ToM.

These terminological considerations are deepened in the paper *The role of metacognition in human social interactions* (Frith, 2012, p. 2214)*: “*Metacognition concerns the processes by which we monitor and control our own cognitive processes. It can also be applied to others, in which case it is known as mentalizing." It is then legitimately questioned whether *implicit* mentalizing should be referred to as a meta-cognitive ability, since the mental states of others that are processed are not *represented as such* (cf. Frith, 2012, p. 2216). Implicit mentalizing mainly involves skills that can be couched in terms of perspective taking and keeping track of the other person's actions. They are furthermore described as "implicit representations" of goals, knowledge and beliefs of others that are generated by adopting the "we-mode" (ibid., pp. 2215–2216).[7] Operating in the we-mode means that the other's perspective is integrated into one's own, thus altering salience and value of the external world by factoring in the other.[8] At a non-accessible, non-controllable level, it is claimed that we adjust our view of the world – even if it has detrimental effects on one's own performance.

While these processes are thought to be fully unconscious, the meta-cognitive process, i.e., explicit mentalizing, is described as a process of conscious monitoring and control that is also reportable. Although Frith claims that this reportability and communicability is what makes human social interactions so unique and thus the human species so special, he reviews empirical findings which suggest that we do not have access to the actual underlying cognitive *processes* that led to a particular *outcome*. Rather, what happens is that people confabulate in order to maintain a coherent explanation of their behavior:[9] This is still compatible with the author's definition of meta-cognition as the ability to generate "reportable knowledge about the processes underlying our behavior" (ibid., p. 2214), since it involves no truth conditions.

Explicit, meta-cognitive skills are thought to be the defining feature that makes humans unique, since they involve reflective awareness. This claim is strengthened by scientists who claim that those brain regions that are thought to be involved in high-level social cognition find no analog in non-human primates (Mars et al., 2012). Explicit mentalizing skills can

[7] A more detailed elaboration on the 'we-mode' can be found in Gallotti & Frith, 2013.

[8] One important case of an ability that is allegedly enabled by adopting the we-mode is joint action. To successfully act in order to reach a certain goal involves taking into account many aspects of the other, ranging from the goal to her estimated motor behavior. In the words of Natalie Sebanz and colleagues (2006, p.70), the leading characters in the research field of joint action: "We propose that successful joint action depends on the abilities (i) to share representations, (ii) to predict actions, and (iii) to integrate predicted effects of own and others' actions."

[9] This is shown in a study in which subjects had to make a choice about which of the faces they were presented with they found most attractive. Afterwards, the experimenter showed them their chosen picture and asked them to justify their decision. However, these pictures were not the ones people had actually chosen and yet they would come up with quite plausible justifications. (Johansson, 2005)

also be described as top-down modulations of implicit processes. The underlying structure is thought to be MPFC, which itself can be divided into functional subregions.[10] A variety of tasks lead to an activation of MPFC, as Frith and Frith (2007) review in their paper *The Social Brain?,* ranging from mentalizing tasks (e.g., false-belief tasks or observation of an interaction), to person perception and subsequent ascription of attitudes to self-perception and attribution of mental states to oneself. What they all have in common is that they involve thinking about either one's mental states (ibid.).

3.2.5. The Mirror Neuron Network

In 1988, Giacomo Rizzolatti and his colleagues found that the discharge of a group of neurons in area F5 in the monkey brain is correlated with whole motor acts rather than single movements. Motor acts, according to the authors, include several movements that form a whole, goal-directed action (in the case of the experiment in question, this action was to grasp food in order to ingest it). An action is defined "as a sequence of movements which, when executed, allows one to reach his goal." (Rizzolatti et al., 1988, p. 503) The scientists classify the neurons according to the motor acts they encode, for example, "grasping with the hand and the mouth neurons" or "holding neurons" (Rizzolatti & Arbib, 1998, p. 188). They furthermore hypothesize that these neurons form a "vocabulary of motor acts and that this vocabulary can be accessed by somatosensory and visual stimuli." (Rizzolatti et al., 1988, p. 491) Different sets of neurons thus represent different motor acts; these neurons are now commonly referred to as 'canonical neurons' and are thought to be the "neural substrate of the mechanism through which object affordances are translated into motor acts." (Rizzolatti & Fabbri-Destro, 2010, p. 223)

In further studies, yet another – and more surprising – property of F5 neurons was revealed: one class of them discharge when the monkey *executes* and action, and also when she *merely observes* the same action (Di Pellegrino et al., 1992). Mirror neurons – as these cells were named – thus possess both visual and motor properties and accordingly fire when presented with visual stimuli of actions and when the monkey executes a motor act. Soon after their discovery, hypotheses about the alleged function of mirror neurons were developed. In

[10] For a review of these details, see Amodio & Frith ,2006.

general, it is assumed that their activity represents actions (Rizzolatti et al., 1996a, p. 138) and they thus serve to understand motor events, which is defined as follows:

> With this term we indicate only the capacity of an individual to recognize the presence of another individual performing an action, to differentiate the observed action from other actions, and to use this information in order to act appropriately.

The discovery of mirror neurons in monkeys provoked the question whether such a system existed in humans as well, which several studies pursued. The data that was gathered from monkey studies that relied on single-cell recordings, which means that the monkeys have implanted electrodes in their heads that stay in their skulls for several days. This method cannot be applied to humans easily, since the only way of doing single-cell studies in human subjects is to combine them with the operational procedure of implanting electrodes in the brains of, for example, epileptic patients.

Fadiga and colleagues (1995) were the first to conduct a study which hinted to a mechanism that matches action production and action observation in humans. They showed that when human subjects observe an action, transcranial magnetic stimulation (TMS) enhances motor-evoked potentials (MEP) that were recorded from the muscles that are usually activated when executing the action. However, this study did not yield any evidence of an underlying neural system with mirror properties. In two further studies, it was assumed that if mirror mechanisms existed in humans, the observation of actions should lead to an activation of both areas with visual properties and areas with motor properties. These experiments, using PET, delivered the insight that the homologue areas to the monkey brain were activated when humans watched the experimenter grasp objects, hence hinting to a mirror neuron system in humans (Grafton et al., 1996; Rizzolatti et al., 1996b). Area F5 in the monkey brain roughly corresponds to Broca's area. They are not only similar in location, but also show cytoarchitectonical similarities.[11]

The core network has been called the *parieto-frontal mirror neuron system* (Rizzolatti & Sinigaglia, 2010) and is composed of two main regions: the inferior parietal lobe (IPL) and the inferior frontal gyrus (IFG). Neurons in these regions show 'mirror' properties, i.e., they fire likewise whether an action is executed or merely observed. Furthermore, these regions

[11] However, the mirror neuron system is not to be found in one single area. In both human and monkeys, it is distributed in neural circuits, including both visual and motor areas. For a detailed description, see Rizzolatti & Craighero, 2004; Rizzolatti & Fabbri-Destro, 2010.

form a circuit with STS, which has no mirror properties. Iacoboni and Dapretto (cf. 2006, p. 946) hypothesize that these three regions functionally work together as follows. While STS provides a higher-order visual description of the observed action, IPL encodes motor aspects and IFG serves to detect the underlying goal of the action.

This core mirror neuron circuit has been hypothesized to be also activated when a person plans (Johnson, 2002), observes (Grafton et al., 1992) or imitates (Buccino et al., 2004) an action. Other studies suggest that mirror neurons fire independently from the visual presentation of the effector, i.e., when the subject is presented with auditory stimuli that are associated with actions usually executed with the hand or mouth (Gazzola, Aziz-Zadeh & Keysers, 2006). Furthermore, the degree of activation seems to depend on prior motor experiences of individuals. Buccino and colleagues (2004) show that motor circuits are activated when subjects observe actions that belong to their own motor repertoire, such as lip smacking or biting in both human as well as non-human (in this case, monkeys and dogs) animals. However, when presented with actions that humans are not capable of (e.g., barking), no activation of motor areas occurs.

A series of studies suggests that the mirror neuron system shows plasticity and individual differences depending on motor skills and experience of subjects. Not only did Calvo-Merino and colleagues (2004) find that motor resonance was greater in the brains of skilled dancers watching dance videos than in lay persons. Cross and colleagues (2006) revealed that activation during the observation of novel skills is increased as dancers learn to execute them. These findings suggest a role for mirror neurons in learning and imitation of motor skills.

Interestingly, Rizzolatti and colleagues argue for two different networks that underlie mechanisms by which social behavior and understanding are achieved. The one is, according to the authors, tightly related to emotional processing and comprises amygdala, orbitofrontal cortex (OFC) and temporal lobe. The mirror neuron system, however, is devoid of emotional components and merely functions to understand the other's actions and motor acts: "The meaning of the observed action does not result from the emotions it evokes, but from a matching of the observed action with the motor activity which occurs when the individual performs the same action." (Rizzolatti et al., 1996a, p. 173)[12] Actions, according to the

[12] The distinction between a system for emotion recognition and action recognition has been spelled out in greater detail by Cattaneo & Rizzolatti, 2009.

authors, gain their meaning in virtue of being represented in cortical areas and associated with their potential outcomes. Mirror neurons are now thought to enable the individual to attribute this knowledge to actions that are executed by others:

> When an external stimulus evokes a neural activity similar to that which, when internally generated, represents a certain action, the meaning of the observed action is recognized because of the similarity between the two representations, the one internally generated during action and that evoked by the stimulus. (Rizzolatti et al., 1996a, p. 173)

The authors point to two questions that follow from this interpretation, both of which are left unanswered. The first asks how exactly the recognition process that is triggered by watching another individual works, that is, by which mechanisms one individual distinguishes her own movements from those of others (cf. ibid.). The second question asks how visual information is associated with motor information. The first question is aimed to be answered by Gallese and Goldman (1998) who apply the finding of mirror neurons to theoretical considerations about a simulation mechanism (see chapter 1.3.5.).

Referring to Jeannerod's (1994) theorizing about the mirror neuron system's functions, Gallese (1996) claims that one plausible interpretation asserts that the mirror neuron's discharge generates internal representations of the observed movement. These representations then occupy several roles, one of which could be motor learning by imitation. The matching mechanism of the mirror neuron system, according to this view, extracts crucial properties of the visually presented stimuli (the observed action) and codes them into motor information, based on the vocabulary of motor acts.

Another function, as already described by Rizzolatti and colleagues, is action understanding, which crucially relies upon the *similarity* between representations of visual stimuli and motor acts. The finding of cells which possess sensorimotor properties also speaks for the common coding theory as described by Prinz (1990). There indeed seems to be a neural mechanism that links perceptual with motor representations, which causes the association of a perceptual event with an action and *vice versa* (Singer, 2012).[13]

[13] Yet another function is assigned to the mirror neuron system, namely to form the basis for communication, speech and language. The basic idea is that the neural mechanism yields a basis for attributing meaning to motor acts, which in turn forms a shared basis for communication. Phylogenetically, it is hypothesized, the precursors of language were *actions,* rather than animal's calls. Before language evolved, understanding gestures and other bodily expressions was even more crucial and can be seen as an early form of communication. Communication – verbal or non-verbal – requires that both receiver and sender share a semantic basis upon which their understanding draws. This shared basis could be 'grounded' in the mirror neuron system which enables individuals to

One of the perhaps most controversial claims about mirror neurons is that they enable humans to understand the underlying *intention* of motor acts, i.e., the motivation and goal of the action. Research on the monkey mirror neuron system suggests such a claim. Umiltà and colleagues (2001), to name one example, showed that neurons in F5 responded to both the observation of a fully visible action and the observation of the same action when the crucial part was occluded. The authors thus conclude:

> This indicates that, even when visual cues are limited, the activation of mirror neurons can place the observer in the same internal state as when actively executing the same action. This would enable the observer to recognize the hidden action. (ibid., p. 161)

Several studies have been conducted to prove this claim that has been extensively criticized (especially by philosophers, see section 7.3.3.). In one such study, subjects were presented with different videos (Iacoboni et al., 2005). The first showed a table with supplies for drinking tea, suggesting to be set up either before or after having it. The second video showed two different ways of picking up a mug, while the third displayed the tea set-up either before tea or after tea with a hand that grasps the mug in different ways (suggesting that the mug is either picked up for drinking tea or for cleaning the table). It was hypothesized that if mirror neurons merely encode motor aspects of observed actions, their activation should not differ depending on the absence or presence of contextual information. This is how Iacoboni and colleagues (ibid., p. 533) interpret their results:

> The present findings strongly suggest that coding the intention associated with the actions of others is based on the activation of a neuronal chain formed by mirror neurons coding the observed motor act and by ''logically related'' mirror neurons coding the motor acts that are most likely to follow the observed one, in a given context. To ascribe an intention is to infer a forthcoming new goal, and this is an operation that the motor system does automatically.

Context plays a crucial role here, since it serves as a predictor for the most likely subsequent action in a given situation. It has thus been concluded that mirror neurons must encode the

understand intentional acts – first, motor acts and, in later course of phylogenetic development also speech (Rizzolatti & Arbib, 1998; Metzinger, 2009). Rizzolatti and Arbib (1998) argue that one piece of evidence for this claim comes from the fact that while the underlying anatomical structures of animal's calls and human speech differ, one part of the mirror neuron system – endowed with representations of both hands and mouth – is found in Broca's area 44, which is thought to subserve language skills in humans. In their words: "Our proposal is that the development of the human lateral speech circuit is a consequence of the fact that the precursor of Broca's area was endowed, before speech appearance, with a mechanism for recognizing actions made by others." (ibid., p. 190) underlying *intention* of the observed action. The role of context for the function of the mirror neuron system will further be scrutinized in chapter 7.3.4.

3.2.6. The Empathy Network

Tania Singer (2012) describes the emergence of research on empathy at the beginning of the 21st century as being motivated by and conceptually grounded in Gallese and Goldman's (1998) seminal paper on ST and mirror neurons, and Preston and de Waal's (2002) suggestion of a neuroscientific model of empathy. Preston and de Waal claim that common coding can be seen as the basic principle to underlie both empathy and action understanding. The principle as described by Prinz and colleagues (1990) suggests a tight link between action and perception because they are thought to share a representational basis. This ultimately means that perception elicits the representations of associated motor acts and *vice versa*. This idea relates to the mirror neuron system as the network supposed to underlie action understanding quite obviously.

Preston and de Waal extend the idea of common coding to the affective realm and claim that empathy follows the same principle: when an emotional state is perceived in another individual, this elicits a representation of the same state and its somatic associations in the observer. Based on Gallese and Goldman's theory, these shared representations can be seen as simulations of sensorimotor, affective or mental states. Thus, the "shared network hypothesis" (Singer & Lamm, 2009, p. 84) as described in the context of social neuroscience holds that the same neural networks are activated both when an individual experiences a particular emotion herself and when she observes someone else displaying this emotion.

While most studies found that the anterior insula (AI) and anterior cingulate cortex (ACC) are especially important for processing vicarious and first-hand emotional response (ibid.), newer findings also suggest a role for somatosensory cortices (Bastiaansen, Thioux & Keysers, 2009). However, several meta-analyses speak against a specific activation of somatosensory areas for empathy and rather suggest that they respond to more general perceptual stimuli of being touched or feeling pain (Keysers, Kaas & Gazzola, 2010; Singer, 2012). There is a bias towards pain studies in the field (Cheng et al., 2007; Decety & Lamm, 2006; Singer et al., 2004), but activation of AI has also been shown for the experience and observation of disgust for odors (Wicker et al., 2003), taste (Jabbi et al., 2008), and touch (Keysers et al., 2004).

Moreover, damage to the insula results in impairments of emotion recognition for disgust, but not other emotions (Calder et al., 2000). The rationale behind the interpretation of this data has been described by Singer and colleagues (2004; 2009) as the interoceptive model of empathy. This model suggests that "cortical re-representations in AI of bodily states may have a dual function" (Singer & Lamm, 2009, p. 86), namely to form subjective representations that are also used to predict bodily effects of specific stimuli and then to form a "visceral correlate of a prospective simulation" (ibid.) that is then used to understand the other person.

Singer and De Vignemont (2006; 2009) provide a carefully thought-out terminology for the field of empathy research. Since the phenomenon is claimed to be related to, but distinct from phenomena such as mimicry, emotional contagion, perspective taking, sympathy and prosocial behavior, this is a great convenience. The definition of empathy that Singer and de Vignemont (2006, p. 435) yield is the following:

> There is empathy if: (i) one is in an affective state; (ii) this state is isomorphic to another person's affective state; (iii) this state is elicited by the observation or imagination of another person's affective state; (iv) one knows that the other person is the source of one's own affective state.

The delineation of other phenomena can be derived from here; while perspective taking lacks the emotional component, sympathy involves no isomorphism, and emotional contagion misses a distinction of self and other. The latter and mimicry are thought to often precede empathy, but to be neither necessary nor sufficient for the occurrence of empathic responses (Singer & Lamm, 2009).

These considerations are important for Singer's and de Vignemont's claim that processes of empathy are crucially modulated by contextual cues at early processing stages. The account they argue against holds that empathy depends on *automatic* activation of emotional, visceral or somatosensory representations when presented with a specific stimulus. Although empathic responses need not to be explicitly *initiated* – this is shown by studies in which participants did not know the goal and were merely told to look at the stimulus, and thus can be called automatic to a certain degree – this does not reflect the phenomenology of empathy. Humans do not – as the authors point out – automatically pick up every emotion they encounter (De Vignemont & Singer, 2006). Thus, Singer and de Vignemont argue for a contextual approach to empathy that considers the manifold top-down effects that modulate empathic responses substantially. Effects such as likeability of the observed person, background and situational knowledge (e.g., justification of emotion), intrinsic features of the emotion that is shared (e.g., intensity, valence and salience) and the relationship between observer and target can influence empathic response at both the behavioral and neural level (for a review of these results, see De Vignemont & Singer, 2006, pp. 437–439; Singer & Lamm, 2009, pp. 88–92). It is thus proposed that "(i) empathy is modulated by appraisal processes and (ii) this modulation is present even at the sub-personal level of a neural empathic response, and can be fast and implicit." (De Vignemont & Singer, 2006, p. 437) Taken together, is it fair to say that while there are exciting findinds for each network that is thought to underlie social cognition, there is just as much uncertainty as to their function. In chapter 7, I will suggest a new way of looking at the social brain.

3.3. Conceptual Challenges

While Ochsner and Lieberman (2001) see social psychology and (cognitive) neuroscience as the main contributors to a meta-theory of SCN, Adolphs (2003, 2010) widens the scope and adds philosophy to an interdisciplinary, theoretical approach to social cognition. He reflects on the emergence of SCN by naming requirements and challenges for the field that are of both methodological and theoretical nature. In this chapter, I wish to highlight two of them:

(a) In chapter 5.1.1., I will focus on the first and probably most pressing challenge, viz., the question of whether social cognition is in any sense special and distinct from general cognition. I present several possibilities to give an answer to that question and show which consequences they carry. I argue that there is indeed reason to assume that there is something special about social cognition.

(b) In chapter 5.1.2., I focus on terminological issues that arise when the claim that social cognition is distinct from general cognition is confirmed and contend that a new set of terms is desirable for an interdisciplinary take on social cognition.

3.3.1. How Special is Social Cognition?

In chapter 1, I listed d_1 – specificity as a desideratum for a theory of social cognition. This is, however, by no means an uncontroversial claim. The question whether there is something special about social cognition is still discussed vividly. In this chapter, I aim at clarifying why I argue for social cognition to be ‘special’. In doing so, I present several positions and possible arguments for and against the specificity claim.

To begin with, it should be emphasized that the answer to the question has important methodological consequences for the interdisciplinary research field of SCN and related domains. Should it turn out to be true that social cognition is in no sense different from non-social cognition or is entirely reducible to it, the target of *social* (cognitive) neuroscience becomes somewhat blurred. It can be argued that the very assumption that the social can be distinguished from the non-social is thus primarily a *methodological assumption* that serves to delineate an area of research. Such an assumption is not too far-fetched, considering that there are indeed plausible arguments for the specificity of social cognition. Stanley and Adolphs (2013, p. 817) elaborate that “[i]n studying the ‘‘social,’’ social neuroscience is about the neurobiology involved in perceiving, thinking about, and behaving toward other people.” However, the questions of what it is that defines social cognition as *social* and what differentiates it from general cognition are yet to be answered.

In order to begin to yield answers, it will be useful to subdivide the specificity question into three further issues. The first concerns *social ontology* and asks whether social (cognitive) processes form an ontologically distinct class of entities. In which sense it is plausible to say that ‘being social’ is a property of cognitive processes and if so, how does that difference manifests itself? Does it lead to an experientially qualitative difference in the same sense in which ‘being conscious’ as a property of a cognitive state alters its experiential nature? If a social event instantiates a particular property at a given time and place, it can further be asked what *kind* of property a *social* property actually is. It could be possible to find a set of minimally defining features of what makes a cognitive process a *social* cognitive process, that is, a set of features that each process which is called ‘social’ must possess. This can also be seen as a set of conditions that need to be fulfilled in order for a specific process to arise. Whether or not we find defining features also influences the search for a demarcation of both the target phenomenon and the research field of SCN.

Another question concerns social epistemology. Here it can be asked whether social processes *require* a distinct and specialized set of epistemic processes and whether social situations expose individuals to an epistemically different situation. Put differently: Is there a different kind of knowledge that is available to the individual in a social context that recruits a different perspective?

These two first questions can probably be most fruitfully tackled by asking a third one: At which level of description are social cognitive processes special and distinct? The quest for a multi-level analysis for a phenomenon as manifold as social cognition has been justified in chapter 1.3.6. and has also been brought forth by the founders of SN and SCN. However, if it is true that social cognition is special, will we find the same degree of specificity at every level of description? One possibility is to assume that if there is something unique at one level of analysis, it should be possible to track this down to lower levels (Adolphs, 2006a, 2006b). Although it does not necessarily follow that behaviors that are *phenomenologically special* recruit specialized mechanisms, the possibility gets more realistic when applied to the broader context of the relation of levels of analysis. Thus what makes social cognition special may lay in the bridges and connections between levels – an idea whose further investigation needs both philosophical as well as empirical contributions.

While it will be difficult to find answers to social ontological questions, there are numerous attempts to clarify epistemological and level-related questions.

Interestingly, both proponents and opponents of a specialized view of social cognition claim that at the experiential and behavioral level, social encounters and interactions indeed are different from dealing non-social ones. Hohwy and Palmer (2014, p. 167), who claim that it makes little sense that (sub-personal) social processes should be seen as specialized, assert that "the perception of things like the intentions and beliefs of other people feels more intangible than, for instance, the perception of visual objects." It seems intuitively right to say that interacting with other people is of high saliency for humans and is also perceived in such a way.

The trickier question, though, is to ask whether this perceptual outcome draws on *domain-specific* or *domain-general* processing:

> Either social cognition just poses a correlated set of computational demands, and is typically associated with a correlated set of effects, none of which are unique to social information; or else social cognition is domain specific, but it is difficult to isolate what the essential feature might be. (Adolphs, 2010a, p. 761)

This quotation nails down the problem that arises at the level of computation and information processing. Adolphs (cf. ibid., p. 753) distinguishes different stages of processing of social information, namely social perception, social cognition and social regulation. While stimulus-driven, bottom-up processes of social *perception* may not be unique in the sense that they recruit – as any other perceptual process – channels for certain sensory processing, this case is different for social *cognition*. Described as a mechanism that functions rather from the top down, it is thought that

> [o]nce we go beyond processing driven primarily by the sensory input and consider the attributions and inferences that we make in order to construct a richer representation of the distal stimuli that caused the sensory input social cognition recruits processes that seem to find no analog in nonsocial cognition. (Adolphs, 2010a, p. 754)

In this interpretation, the social cognitive processes themselves are viewed as unique and not to be found in more general cognition. However, to understand what is special about social cognition according to the author, a closer look is needed.

It is claimed that social cognition can be equated with the sum of mechanisms that enable humans to perceive aspects of the world as socially meaningful (Adolphs, 2010a). Thusly characterized, its defining feature would be a *functional* and *epistemological* one. It is furthermore assumed that *meaning* is only accessible through an inferential mechanism, which in the case of grasping *social* meaning is prominently referred to as ToM, mindreading or mentalizing. What is hidden and needs to be inferred are the mental states of other people. The essence of what differentiates social cognition is not merely that it recruits inferential mechanisms and depends on filling-in psychological states. As Adolphs acknowledges, *all* perception involves filling-in of those aspects that are not perceivable from one's perspective. It is the very fact that *mental* or *psychological* states need to be inferred. This again boils down to the claim that it is the *kind of information* which makes a cognitive process a special, social one, or not.

This leads to yet another candidate for what makes social cognition special, which is often only implicitly stated when researchers try to define the phenomenon. Social cognition is frequently referred to as an ability which enables individuals to navigate their social world, thus implicitly assigning a specific *function* to it. Although this is a truly broad definition, it makes sense from a variety of perspectives. It restricts the set of processes and mechanisms that are under scrutiny to those that serve individuals to fulfill social tasks, thus yielding a possible way to demarcate the target phenomenon for SCN.

A potential counter-argument against the claim that social cognition is unique in the way suggested above is described by Adolphs (ibid., p.754) in the following manner:

> Of course, one can argue with this view of social cognition as special and see the story instead much like what I sketched for social perception: inference, attribution, and filling-in are of a special sort for social stimuli, but the general computational process is ubiquitous in how the brain processes information.

This is exactly what Hohwy and Palmer (2014) argue for in their paper *Social cognition as causal inference.* In their view, social cognition is a representational affair that is referred to as 'mentalizing' and although social sensory *input* is unique in many respects, the underlying *mechanism* is not. Neither is there a difference at the representational level: "The representation that occurs in mentalising is entirely analogous to the representation that occurs in non-social contexts." (ibid., p. 169) The difference lies in the *challenge* a certain class of stimuli poses to the computational mechanism, but not in the mechanism itself. How different stimuli are processed remains the same, no matter the variation in the complexity of computational demands a stimulus evokes. Hence,

> [i]f words, gestures, and additional behaviours that we pick up from other people are treated as just being characteristics of sensory input, and if the mental states of other people are treated as the causes of this input, then mentalising can be characterised as causal inference from sensory effects to worldly causes […]. (ibid.)

The authors arrive at this characterization by naming different candidates of how social cognition could be thought of as special. Although they acknowledge the complex nature of the social world individuals have to handle and the high level of uncertainty that comes with this complexity, this is not thought to be entirely unique to the social realm. Context-dependence and a seemingly intangible amount of causes can be ascribed to explaining the behavior of an individual as well as to the financial crisis, to pick up an example of the authors. However, the difficulty of inference of behavioral causes stems from the sparse number of *sources of evidence* in mentalizing. Basically, observable behavior is the main source that inferential processes can draw on, making it especially difficult. A low amount of evidential sources is not restricted to the behavior of individuals, but since evidential insulation occurs *systematically* in mentalizing, it is acknowledged as a marker for social cognition.

So far, what the authors offer is a proposal of what makes social cognition special, viz., the kind of input it processes. However, there is something dubious about Hohwy and Palmers line of argument here. Their view is only valid if one agrees with the claim that in social situations, there is a systematic shortcoming of evidential sources. That description seems to profoundly undermine the manifoldness of stimuli in a social interaction. Folk psychologically speaking, we surely refer to 'behavior' as if it was unitary, but a closer look easily reveals that this is not the case at all. Behavior can be conceived of as being composed of different sources that can build a rich basis for inference. Examples are facial expressions, gestures, body posture, etc. The claim that behavior is no unitary phenomenon can be strengthened by the fact that most research on social cognition is not about ‘behavior’, but about exactly those aspects.

I argue, furthermore, that behavior is not the mere basis for inference. There are ways to check on one’s interpretation of behavior, such as asking the person why she did something, making use of prior knowledge, consulting other people, increasing one’s general knowledge about human behavior, etc. Also, as has been proved empirically in the past decades, being exposed to social situations elicits (bodily as well as neural) responses in all participants that can serve as a further source. Thus, it is doubtful that their evidential sparseness is what makes social stimuli distinct.

One possible hypothesis at this point is the following. Even though there is a manifoldness of stimuli that can serve to disambiguate and understand another individual, there is still something that is unavailable and not mediated by these stimuli. This would ultimately mean that there indeed is something hidden about minds that does not necessarily manifest itself in the kind of cues that are available to other people.

Summarizing Hohwy and Palmer’s view, there are aspects of the social realm that differ from the nonsocial, but this merely refers to the *degree of uncertainty* that is to be found in social stimuli. *Computationally,* these stimuli are processed just as any other sensory input:

> Social cognition, we therefore propose, is nothing but causal, Bayesian inference from sensory input to mental states. To understand social cognition and how it may differ from other areas of cognition, the task is then to specify how uncertainty may arise in the inference from sensory input to mental causes. (Hohwy & Palmer, 2014, p. 170)

Besides the agreement of Adolphs and Hohwy that sociality is special at the macro-level, both authors see convincing evidence for social cognition drawing on (at least partly) domain specific *neural* mechanisms. As described above, a modern version of SBH does indeed identify several networks that correlate with social behavior and cognition (see chapter 3.2.1.). The basic question, then, is whether these networks are specialized or have evolved for processing social stimuli or whether “all social cognition draws on entirely domain-general processes, only applied to social stimuli.” (Stanley & Adolphs, 2013, p. 822)

As I already showed, the idea of social domain specificity tightly relates to the idea of modularity. In the book chapter *What Is Special about Social Cognition,* Adolphs (2006b) lists features of modularity from both a Fodorian view, as well as from more current perspectives and asserts that social cognition as the *sum* of many social mechanisms draws on too many domain general processes to be called modular in a strong Fodorian sense. However, there are indeed *single aspects* of social cognition where empirical evidence strongly suggests domain specificity. Specialization, though, is not the same as modularity, although domain-specificity is one important feature of modules. Thus, the claim that social cognition is special in one sense or the other still holds if modularity is rejected.

In summary, I have presented several candidates that could be seen as defining the 'social' in social cognition. Let me now come back to the three questions I asked at the beginning of this chapter. I argue that all of the views I just presented – even the ones that explicitly deny specificity of social cognition – can be interpreted as agreeing on one assumption which applies to all three questions. This is that in a functional sense, social cognition is special.

In tackling the social ontology issue, it can be argued that what makes a property of cognitive states a social one is their function. Functional properties of cognitive processes enable an individual to socially interact with its environment. Furthermore, these mechanisms process social stimuli, which have been classified as uniquely complex. Although Hohwy and Palmer disagree with a specificity view on social cognition, their perspective can still be seen as agreeing with functional specificity. This is because they claim that social stimuli are distinct from non-social stimuli. These specific stimuli must be processed by mechanisms that function to disambiguate them and in this sense it can be said that social cognition is special.

This also relates to the social epistemology question in the sense that the epistemic processes needed for social cognition can be individuated by their functional properties. Moreover, the third, level-related question can be answered by stating that at the functional level of description, the specificity claim holds. In chapter 7, I will further elaborate on how this functional specificity relates to other levels of description.

3.3.2. Terminology

In chapter 1.5.1., I argued that the current set of terms which is available for the philosophical discussion on social cognition is rather confusing and is thus in need of renewal. This terminological problem, it seems, is not only one of philosophers.

Among other things, the problem of whether or not there is something special about social cognition also provokes terminological issues. This point is nicely summarized by Adolphs (2003a, p. 121):

> It is clear, on the one hand, that our behavior is in many respects specialized for guiding our interactions with others, that social behavior is more complex than other aspects of behavior, and that our social behavior is at least quantitatively (and perhaps qualitatively) vastly different from the social behavior of other species. On the other hand, it is equally apparent that there is phylogenetic continuity in both brain and behavior, and that social cognition depends on many of the same mental processes, and hence presumably on many of the same brain structures, as do non-social aspects of cognition. Do we need an additional vocabulary to explain the cognitive processes that guide social behavior? Or can social behavior be linked to neurobiology using existing, domain-general constructs, such as attention, memory, and so forth?

In other words, if social cognition is indeed special, does this lead to the need of a special set of terms to describe the neural, cognitive, and behavioral aspects of the phenomenon? Research needs a proper terminology to successfully describe its target phenomena. The need for such a terminology relates to one question that Adolphs (2003a, p.124) poses, namely if SCN should draw on already existing concepts or rather needs a “proprietary vocabulary”. While social psychological terms such as ‘attitude’ or ‘attractiveness’ express domain-specific social content, neuroscientific language operates on terms like ‘memory’ or ‘language’, referring to domain-general processes. This leads to a dichotomy that is rather hard to reconcile and shows that the interdisciplinarity of the research field poses a specific problem to a terminology. Since, as shown above, different areas use different concepts and terms, the quest for a *common language* has occurred.

The issue of a common terminology that is familiar to both social psychologists and cognitive neuroscientists is brought up by Ochsner and Lieberman (cf. 2001, p. 719), who emphasize that drawing on a shared pool of terms facilitates communication between researchers. The question is, however, how realistic the endeavor of finding a shared terminology is. In a recent paper, Stanley and Adolphs suggest that a first step towards finding a set of terms is to accept that there are different ones at each level of description that not necessarily translate into each other or for which we might now yet have plausible bridging principles. The key to success is described as follows:

> […] reduction or elimination is not needed: what is needed is communication, so that those working at different levels of analysis can appreciate, and understand, work at different levels. We do not so much need a single language, as we need people who can speak several languages and translate easily between them. (Stanley & Adolphs, 2013, p. 822)

While this sounds like a more achievable goal for a broad field like SCN and its related disciplines, some problems remain.

Even if we do not need entirely unitary terms in the sense that each discipline describes the same phenomenon from different perspectives and thereby uses a different set of descriptions, the meaning of these descriptions should be agreed upon by the participants of the research process. In chapter 1.5.1., I showed that this has not worked for central concepts such as ToM, mentalizing and mindreading. It was argued that the extensive usage of these terms in a vast amount of different contexts has stretched them too far and thus left them semantically vacuous. I therefore propose that we should pay attention that those terms that are being used are not 'exploited' like this and keep their expressiveness. This does not mean, however, that a 'metaphorical' use of terms should be prohibited. Much rather, I contend that it is of great importance that those terms that are being used have a meaning that is widely agreed on. This is not given for the concepts being used in the mindreading debate, which is why I suggest that a new set of conceptual tools will prove fruitful for the interdisciplinary discussion of social cognition. The desideratum 'terminological consistency' thus also relates to the dialogue between disciplines in the sense that our terms at hand should be used by researchers from various disciplines in a consistent manner.

The question of translatability not only concerns the dialogue between disciplines. It has also been asked whether it is worth striving for to couch neuroscientific findings into folk psychological and thus more understandable terms (Adolphs, 2003a). Cognitive neuroscience, according to Adolphs, will most probably be able to show which and how brain mechanisms constitute social processing. However, in order to make them valid and meaningful explanations, a sound theoretical background is needed that possesses a vocabulary which enables to translate the one into the other. If SCN will not be able to relate folk psychological concepts to its empirical findings, it will fail to provide meaningful explanations of social behavior. Thus, one main desideratum for a future language is to bridge folk conceptions with social psychological and cognitive neuroscientific explanations. In doing so, it might not be sufficient to alter already existing concepts, but to "invent a new set of terms that can translate between the different ways of describing social behaviour, and that correspond more closely to the neural processes that underlie them." (Adolphs, 2003c, p. 176)

A thorough philosophical analysis of already existing concepts can, in my view, provide a more detailed perspective on which terms are useful and which have considerable shortcomings. Thus, a philosophically well-elaborated meta-conceptual taxonomy of terminological desiderata could suggest ways in which we can begin to find an appropriate and fruitful terminology for SCN and its related disciplines. This was my aim in this and previous chapters; to suggest ways in which we may begin to form a new set of terms will be one main goal in Part II of this thesis.

Interlude

4. Interlude: The State of the Debate on Social Cognition

> *To conclude our diagnosis: neither COG [classic cognitivism] nor ENAC [enactive cognition] has been successful in providing a convincing account of both online and offline forms of cognitive processing. It hence seems fruitful to aim at a unified theoretical framework that solves the stalemate between ENAC and COG and integrates online and offline processes into a coherent story of how cognition can best be understood. (De Bruin & Kästner, 2012, p. 547)*

What De Bruin and Kästner express in this citation is that twofold. First, it shows that there is a spectrum of theoretical claims, whose ends appear to be classic cognitivism and enactivism. Further, they rightly diagnose that either account is yet to come up with a comprehensive and coherent account of cognition. The goal of this chapter is to show that the same holds for social cognition. In doing so, I draw on the contributions of each field that have been evaluated so far, that is, philosophy of mind, phenactivism, and social cognitive neuroscience.

I argue that in order to find a theoretical framework that is able to capture the multitude of social cognition, we need to find a way to integrate insights from both ends of the theoretical spectrum. Such a comprehensive view does not come, as will be shown, without challenges. Having sensed the shortcomings of mindreading theories of social cognition, philosophers and scientists have lately suggested to turn towards enactive accounts of cognition and social understanding. In chapter 4.1., I will describe these movements that have been dubbed ‘the pragmatic turn’ and ‘the interactive turn’. I further examine this development and argue that while integrating ideas from phenactive views on social cognition will prove fruitful, a full turn is neither necessary nor desirable in the attempt to from a consistent theoretical framework on the phenomenon. I will then scrutinize a fairly recent theoretical approach to social cognition whose basic thought is extremely useful for our endeavor. These so-called pluralistic accounts of social cognition aim to combine elements from both ends of the theoretical spectrum. This combination, as I argue, does not come easy and needs to be careful not to end up putting together metaphysically incompatible assumptions.

4.1. A Shift in the (Social) Cognitive Sciences?

This subchapter deals with current developments, namely the pragmatic turn in cognitive science and the interactive turn in the research field on social cognition. While I will only briefly touch on the pragmatic turn, I focus on the interactive turn in SCN and related disciplines. In short, proponents of these turns claim that research is shifting towards enactive paradigms of (social) cognition and are thus turning away from cognitivist views.

The central questions I will ask in this chapter are the following: To which degree is the interactive turn *actually* happening and to which degree *should* it happen in order to be a fruitful development? Further, if we assume that the research field should shift its focus in order to capture hitherto neglected aspects of (social) cognition, another question arises. Does the *methodological* shift that is needed to investigate those aspects *necessarily* involve a *theoretical* shift as well?

To answer these questions, I will take three steps. First, I aim to show that while the interactive turn indeed takes place to some degree, it is by no means an ultimate turn towards the phenactive end of the spectrum of paradigms. I hypothesize that researchers are reluctant to fully commit to phenactive theories for several reasons. I argue that the turn is thus better described as an educational journey with a re-turn. Secondly, I claim that both the turn and re-turn make sense, since no radical position might be suitable for a theory of social cognition. This leaves us with the need for a paradigm that is able to integrate insights from both ends of the spectrum. In a third step, I investigate which ideas of the interactive turn should be adopted in order to reach the goal to provide an integrative and comprehensive theoretical framework for social cognition.

The chapter will proceed as follows:

(a) In chapter 4.4.1., I exemplify the pragmatic turn in cognitive science referring to Engel and colleague's (2013) work. I show that while they lay out a fairly radical theoretical shift, they still hesitate to fully commit to their claims.

(b) Chapter 4.4.2. describes the interactive turn in SCN and focuses on a high-impact paper by Schilbach and colleagues (2013), which aims to lay out a theoretical and empirical foundation for a throughout shift in the research field.

(c) In part 4.4.3., I critically appraise the claims that have been depicted and proceed to argue for the three claims as described above.

4.1.1. The Pragmatic Turn

In previous chapters, I have already depicted the assumptions of the phenactive view of cognition (2.3.1.) and social cognition (2.3.2). The claim of proponents of the pragmatic turn is that the whole research field of cognitive science turns towards those phenactive theories and also adopts the implications that come with that move. In the words of Andreas Engel and his colleagues (2013, p. 202):

> In cognitive science, we are currently witnessing a 'pragmatic turn', away from the traditional representation-centered framework towards a paradigm that focuses on understanding cognition as 'enactive', as skillful activity that involves ongoing interaction with the external world. The key premise of this view is that cognition should not be understood as providing models of the world, but as subserving action and being grounded in sensorimotor coupling.

This includes theoretical, empirical and conceptual consequences. The paper is a good example of researchers being attracted to alternative, i.e., non-cognitivist theories of cognition. At the same time, it also nicely shows that only few members of the research community are ready to commit to those views in all their breadth. After having outlined the theoretical assumptions of the pragmatic turn,[1] Engel and colleagues claim that there is plenty of evidence showing that cognition is much more plausibly described in phenactive terms and that even already existing evidence is better accounted for when reformulated within this scheme. This seems to be the point, as will also become clear for the interactive turn, at which a full commitment to phenactivism fails. While the authors first claim that cognition should not be viewed as entailing models of the world (see quotation above) and reject representationalism, they go on using the jargon of cognitivism to describe empirical findings.

Claiming that sensory processing is 'action-related', the authors describe the role of motor circuits for cognitive processes. In doing so, they name several examples in which the prediction of the outcome of an action influences other cognitive processes (e.g., modulation of attention). While this clearly is evidence for the action-*relatedness* of cognitive processing, it remains unclear how this speaks for a *phenactive* view. The authors themselves explain the data they mention in representationalist terms, drawing, for example, explicitly on Wolpert and colleagues (2003) and their theory of forward *models*.

[1] These include the rejection of representationalist assumptions of the cognitivist paradigm, which I will not repeat at this point.

Moreover, the data presented is compatible with any theory that acknowledges the fact that action changes neural activity (why would it not – any bodily input changes neural activity). However, this rather uncontroversial claim can be accounted for within any non-phenactive theory. The enterprise to show that action modifies *neural a*ctivity and to label this modification as a change in sensory processing again lies the focus on the brain, instead of truly adopting the view that cognition is inherently relational and does not start in the brain and then only 'leaks out' into the world.

A full pragmatic turn, as I see it, would have to come with more than the assumption that cognitive and sensory processing are dependent upon or related to action. The authors seem to come to the same conclusion in the beginning of their paper, but then fail to fully commit to their insights. What they exhibit, after all, is the conviction that cognition is *modulated* by action, but not that action *is* cognition, as they claim at some point to be the central premise of the paper (cf. Engel et al., 2013, p. 203). The full commitment to phenactivism and its assumptions, however, would be essential because it is more than mere harping on about principles. Phenactivism grew out of the deep conviction that cognitive science is doing something fundamentally wrong in assuming the location of cognition to be skull-bound. Full-fledged phenactivism is serious about the rejection of cognitivism and representationalism, because it holds essentially different background assumptions (see chapter 2.3.1.). Thus, the pragmatic turn – if taken seriously – is only complete if these assumptions are acknowledged. For anything else, there are already views of extended, embodied or embedded cognition that mostly come without the rejection of the basis of cognitivism (Rowlands, 2009).

4.1.2. The Interactive Turn

I will now turn to the focus of this subchapter – the interactive turn in SCN and its related disciplines. Equally to the pragmatic turn, this movement is proposed to be a paradigm shift towards theoretical, empirical, conceptual and methodological claims of phenactivism (cf. Gallotti & Frith, 2013, p. 160; Overgaard & Michael, 2013, p. 4). One central assumption that motivates researchers to take the interactive turn is that social cognitive processes are fundamentally different when executed and used in interactive (as opposed to 'observational', non-interactive) contexts. Based upon this assumption, it has been claimed that interaction is irreducible to individual social cognitive processes and hence needs to be considered a proper level of analysis (cf. De Jaegher & Di Paolo, 2007, p. 491), and that being immersed in an interaction changes both quality and quantity of the information available for each individual (cf. Schilbach et al., 2013, chapter 2). Przyrembel and colleagues (2012) suggest that this specific content changes the perspective, i.e., the epistemic viewpoint from which information is processed, from a first-, or third-person perspective to a second-person perspective.

These claims have interesting consequences for both theoretical as well as empirical research of social cognition. As for experimental designs, this new focus asks for the implementation of set-ups with two or more individuals who interact with each other. This can be seen as parallel to the demand in cognitive science to shift towards more ecologically valid environments in laboratories. Theoretically, it is claimed that since 'classical' theories (i.e., ST and TT) have focused on internal, individual processes that take place in non-interactive contexts without emotional engagement and thus neglect the alleged most important part of social cognition, they have failed to yield an exhaustive account of the phenomenon. As De Jaegher (cf. 2009, p. 535) puts it, the interactive turn shifts the problem space of social cognition to social interaction. Similarly to the pragmatic turn, taking the interactive turn comes down to adopting phenactive theories of social cognition and thus to shifting the focus in a non-trivial way. Again, the necessity of this shift is born out of the conviction that social cognition is essentially different than cognitivist views assume.

Although one can observe an increasing number of researchers acknowledging that interaction, emotional engagement and embodiment are important factors, there is still reluctance to fully adopt a phenactivist perspective. One clear example is Schilbach and colleague's (2013) work which aims to lay a foundation for a new research paradigm in social neuroscience. By moving "[t]oward a second-person neuroscience" the authors wish to "throw light on the dark matter of social neuroscience" and to support the research field to "really go social" (ibid., p. 393). This includes the rejection of 'spectator theories of other minds', i.e., ST, TT and other theories that do not acknowledge the (alleged) interactive and emotional nature of social cognition. Schilbach and colleagues propose a theory which depicts the phenomenon as basically drawing on social interaction and emotional engagement. These are seen as "constituents of a second person approach" (ibid., p. 396). This proposal is not only supposed to be embedded in general phenactive theories of cognition, but also on the more specific ones of social cognition which I have described in chapter 2.3.2. What exactly are the claims and arguments?

The central assumption is that social cognition and the processes it recruits are fundamentally different when occurring in an interactive and emotionally laden situation than in an observational and emotionally rather detached situation. This entails a difference in the *kind of 'grasp'* of the other person and amounts to a qualitatively and epistemically different kind of access to the other person. A 'second-person' approach aims to capture these aspects and in that way intends to correct the focus towards the 'truly social' aspects of social cognition. It is claimed that being emotionally immersed in an interaction not only alters the phenomenology of a situation, but also recruits a different set of neural mechanisms and modulates them profoundly. Emotions could thus facilitate "more cognitive ways of understanding minds" (or their absence could make them harder) (Schilbach et al., 2013, p. 397). The authors claim that the role of emotional engagement for social cognition has systematically been underestimated and neglected (see below for a discussion of this accusation). The same holds for social interaction.

Again, it should be stressed that the claim here asks for a non-trivial inclusion of interaction into *both* theory and empirical testing of social cognition. Referring to De Jaegher and Di Paolo (2007), the authors support the view that in some cases, interaction *constitutes* social cognition. There are three aspects of this claim that, according to Schilbach and colleagues, have an impact on research. First, there are two possible roles for an individual in an interaction (initiator or responder), that also have different neural correlates. Secondly, because motivations and intentions are not only shared, but also newly created within the dynamics of the interaction, the outcome and success of the interaction are altered. Since these processes are fundamentally different than those involved in observation, it is argued that they influence social cognitive and neural mechanisms. Thirdly, interactions always need to be considered as coming with a history that needs to be taken into account.

These assumptions are thought to yield the theoretical foundation for second-person neuroscience. The authors wish to show that evidence from developmental, social and cognitive psychology already proves the developmental as well as general primacy of emotional engagement and interaction for social cognition (cf. Schilbach et al., 2013, pp. 397–399). Findings in developmental psychology are taken to put constraints on research of social cognition, since they are interpreted to show that the earliest social skills arise within emotional and interactive contexts. Thus, they need to be taken seriously and implemented methodologically.

This includes to agree on the "second-person argument" which assumes that "an appropriate development of awareness of other minds depends on the infant first experiencing minds which are directed toward her" (Schilbach et al., 2013, p. 397) and on the background assumption that "cognition is grounded in basic perception and action processes and emerges out of the interaction of the organism with its environment" which has the empirical consequence "that – rather than treating it as an experimental confound – a social context and social interaction can be treated as an independent variable of experimentation." (ibid., p. 398)

Such a view also implies rejecting a mechanistic understanding of interaction, i.e., rejecting the assumption that all parts of the interaction can be traced back to individual mechanisms and then be subdivided into ever smaller parts. That is so, as already said, because the interaction dynamics themselves emerge as an irreducible pattern and thus form a new "vehicle for the acquisition of knowledge" (ibid.). The next step to be taken is to apply these convictions onto the methodology and theory of social neuroscience and thus to help the research field to truly go social.

The authors proceed to set up their research program for second-person neuroscience by claiming that their own paradigms already yield preliminary evidence for their claims. In several studies it has been investigated that neural activation changes depending on whether a person is personally addressed, i.e., directly gazed at, by an anthropomorphic character she is presented with in the scanner (cf. ibid., pp. 400–402). Being gazed at is thought to involve emotional engagement:

> When we are personally addressed by others, the perception of their mimic behavior relies, in neurobiological terms, upon tight perception-action coupling with affective and body-based processing feeding into and promoting the preparation of motor responses as a way of picking up and responding to the possibilities for interaction. (ibid., p. 402)

On grounds of the evidence from psychology and this preliminary evidence,[2] the authors put forth their claim that a second-person neuroscience is needed.

4.1.3. A Reluctant Turn

In the following, I claim that there is a reluctance to fully commit to the interactive turn and examine where the reluctance that may prevent a full turn comes from. In general, I want to

[2] See Schilbach et al., 2013, pp. 399–406 for further evidence.

go back to my claim that enactivism comes with strong background assumptions and its full acceptance comes with rather extreme consequences. I start with critically assessing the second-person approach to social cognition as another example of reluctance to fully commit to enactive claims. The next step to be taken is to emphasize that there is no such thing as ‘weak phenactivism’ in a sense that would inform the discussion about research on social cognition in meaningful ways.[3] Or, put differently, the way in which some researchers like to talk about ‘weak enactivism’ has little to do with phenactivism itself and rather depicts some sort of externalist position. Used in that way, it rather causes confusion and blurs the distinction between theories that hold very different assumptions about the metaphysical nature of (social) cognition. Thus I vote for reducing confusion and using phenactivism for the kind of theories that exhibit the assumptions as outlined in chapter 2.3. I agree with Gallagher (2013, p. 422) when he claims that

> [t]he enactive interpretation is not simply a reinterpretation of what happens extra-neurally, out in the intersubjective world of action where we anticipate and respond to social affordances. More than this, it suggests a different way of conceiving brain function, specifically in nonrepresentational, integrative and dynamical terms […]

This quotation nicely shows just how radically different enactivism and phenactivism are from moderate external views. Since I have already laid out enactivism and what it rejects, I here want to focus on the consequences of adopting a phenactive stance and on arguing why we have good reasons be reluctant to accept them.

To start, I wish to put to question in how far the empirical paradigms that Schilbach and colleages present as yielding evidence for phenactive claims really succeed in revealing strong roles for interaction and emotion and thus are able to substantiate second-person neuroscience. To begin with, the ecological validity of the empirical designs can be called into question. In the experiment described in the previous section, the person lies in the scanner, thus no movement is possible. She is presented with a virtual character that is controlled externally. This is neither close to any real-life situation, nor does it necessarily involve emotional engagement. Furthermore, let us assume for now that interaction dynamics do emerge between individuals and then form a constitutive part of the process.

[3] Aizawa (2010) has coined the term weak enactivism to refer to a version of Alva Noe’s sensorimotor theory. This weak claim states that for perceptual experience to arise, one must possess sensorimotor knowledge. The strong claim is contrasted and states that additionally to the possession of this specific kind of knowledge, one must also e*xercise* it (cf. ibid., 2010, pp. 263–264). This has little to do with the current work.

How would this be possible if one of the interacting individuals is an externally controlled avatar? Given these circumstances, could interaction patterns still spontaneously emerge as predicted by the theories which Schilbach and colleagues claim to foster?

Another reason to be skeptical about the validity of their evidence in favor of phenactive claims is that they do not succeed in couching their results in terms of the theory. While the research community applauds Schilbach and colleagues (2013, pp. 414-441) for the general idea to pay closer attention to interaction and emotional engagement in social cognition research, many commentators identify problems with how the authors apply their theoretical foundation onto empirical work. As Gallagher and colleagues (cf. 2013, p. 422) correctly diagnose, the authors do not fully commit to the theory they put forth and remain on computational, cognitivist territory, both conceptually and by keeping the idea of neural correlates of social interaction. Indeed, the authors keep using cognitivist vocabulary to account for the findings they review. By that they seem in no way to get rid of views of cognition that focus on internal mechanisms that are imputed to the brain.

This is problematic for several reasons. First, it triggers the question whether the research field actually *needs* a different theory that comes with a different terminology – or whether it is already well enough furnished. Even though I came to the conclusion that the terminology at hand in SCN and related disciplines is not without flaws (e.g., because the extensive usage of terms in different contexts leaves them semantically vacuous, see section 1.5.1.) and that the theoretical landscape is in need of improvement, this does not necessarily mean that a full substitute of *all* the assumptions that are held is necessary. I will come back to that point shortly. Secondly, the author's falling back into 'old habits' may be symptomatic of the fact that 'throwing out the baby with the bathwater' often ends up being a barrier for progress. Both sides of the theoretical spectrum – i.e., phenactivism and cognitivism – probably have their advantages, their fruitful approaches, and may shed more light on some phenomena better than the other theory. However, this very circumstance should rather motivate to 'cherry pick' from both theoretical strands and pave the way to a comprehensive theoretical framework, instead of dismissing one side entirely and thus neglect possibly crucial insights.

Indeed, this seems to be what is happening in the research field. Although some argue that the interactive turn will end up – once finished – in a full commitment to phenactive theories, the examples I gave above rather point to a middle-way.

Of course, ideas of the interactive turn are causing a shift in research on social cognition. They motivate new research directions and the consideration of alternative theories that could possibly substantiate them. However, there is also a strong reluctance to fully commit to phenactive theories and the empirical consequences. In the next section, I argue that this reluctance is justified and one can observe a 're-turn' to less radical theoretical grounds. I therefore suggest that the interactive turn should be seen as an 'educational journey' that has enriched the research field in new and exciting ways, but that also takes researchers farther than necessary.

One of the consequences of the interactive turn is a profound shift of the *unit of analysis* for SCN, i.e., of the explanatory target for research. Gallagher and colleagues (2013, p. 422) state that the full commitment and adoption of enactive theories changes the research target from individuals (or parts of individuals, such as their brains) and how they interact to the *relation between those individuals itself*:

> The explanatory unit of social interaction is not the brain, or even two (or more) brains, but a dynamic relation between organisms, which include brains, but also their own structural features that enable specific perception-action loops involving social and physical environments, which in turn effect statistical regularities that shape the structure of the nervous system […].

This is due to the assumption that the 'in between' itself has to be considered a structure that holds epistemic value and causal influence for the cognitive process in question. What adopting a phenactive position thus boils down to is not to take a radical *externalist* position, but a *relationalist* one that depicts the non-localizable 'in between' as causally effective.

Another similar consequence is the change of *levels of analysis*. Not only would the approval of interaction as a structure with a proper ontological status indeed demand to make it a distinct level of analysis. Also, the rejection of representationalism and functionalism would eliminate, or at least fundamentally alter, a widely accepted model of levels of analysis that includes the representational level of description. One reason for the reluctance to adopt enactivism could be that it is unclear how those levels would be substituted and, more generally, how enactivism could account for 'representation-hungry' processes, such as dreaming, mind-wandering, or theorizing about another person.

One important question that should be considered in the process of turning away from cognitivist paradigms is why we would want to adopt phenactivism in the first place. What would be advantages and at which point does it yield better or more plausible explanations or 'epistemological tools' than non-phenactive theories? This question is still open and up to both theoretical and empirical research, but as for now, there are only few points that *only* phenactivism can account for. Phenactivism is only needed, or so I argue, if one wants to make the strong claim that (sometimes) interactive patterns constitute social cognitive processes in the sense that no individual process alone – even the sum of two individual processes – could bring about the cognitive process in question. To make this claim, one indeed needs the assumption that interaction is a structure that has its proper metaphysical status. The rejection of this claim and the weaker assumption that interaction is the product, for example, of the actions of two individuals that are directed at each other, could not account for the position. However, even phenactivists or those who claim to support this theory often seem to confuse enabling and constitutive conditions, as I showed in chapter 2.3.5. It is well possible that non-phenactive theories can account for interaction and emotions as enabling social cognitive processes and thus no strong phenactive background is needed.

Other than that, no non-phenactive theory is committed to reject other main claims of the theory, such as the developmental and general primacy of interaction, the importance of emotional engagement, and the need for more ecologically valid research paradigms. This triggers the question of how real the 'enemy' actually is – do 'classical' theories fall as short in explaining social cognition as they are accused to do? Overgaard and Michael (2013), for example, nicely show that both ST and TT are in no ways committed to reject central claims of interactionist or phenactive theories and would even support most of them.

Given that SCN arose out of the interest in emotions and the role they play for human (social) cognition, the accusation that the issue has been neglected systematically appears rather strong. There is plenty of work that relates emotional to social processing. To be fair, though, there seems to be a crucial difference in how SCN and proponents of the interactive turn treat the case of emotions. While the latter focus on emotional *engagement,* the former have concentrated on an individual's emotions and the effects they have on decision-making or social judgments (Adolphs & Anderson, 2013). The claim of a second-person approach is aimed to be fundamentally different in that it suggests that (among other things), emotional engagement in an interaction alters the situation profoundly.[4] While detached observation

[4] "The primacy of second-person engagements creates serious conceptual and methodological problems for psychological research: It demands that emotion be taken as central to an awareness of minds and focuses on emotional responses rather than reflections or constructs." (Schilbach et al., 2013, p. 398)

would require a third-person grasp, it is argued that in an interaction, knowledge is gathered by taking a second-person perspective (Przyrembel et al., 2012; Schilbach et al., 2013, p. 397). Thus it could be argued that the role of emotional engagement is stronger and more central here than in accounts that aim to measure the effects and role of emotions for first-, or third-person knowledge. Both takes on the role of emotions make clear that one more component should be added to the list:

c_9 – emotions
Social cognitive processes are often influenced by and influence emotional processes.

I chose the word ‘emotions’ instead of ‘emotional engagement’ in order not to exclude emotional processes that do influence social cognition, but are not necessarily related to engagement.

Considering the role of interaction and emotions is supposed to make the research field ‘truly social’. However, there are two questions that should be asked in response. First, in light of SCN’s interdisciplinary history (see chapter 3.1.), it is questionable whether the research field is not yet ‘social enough’. Social psychology, for example, starts with the assumption that interactive situations present individuals with fundamentally different challenges than non-interactive contexts. By making the ‘social’ a proper level of analysis, SCN assigns an important role to it as well. We should thus seriously reconsider which elements research lacks and how ‘truly social’ SCN already might be.

Another valid question at this point is also whether or not the interactive turn is actually happening to the extent that is described by its proponents. One aspect that speaks against a ‘full’ turn is that a good part of research remains on cognitivist, representational territory. Although the importance of considering interaction, the body and emotions is not necessarily denied and the need for new methods to shed light onto those matters is well respected, a radical shift is not in sight. The reason might be that it is questionable whether a methodological shift necessarily involves a theoretical shift. It is possible that non-phenactive theories explain the results of more interactive or emotionally salient empirical studies just as well. Indeed, even Schilbach and colleagues, who put a lot of effort into laying out their theoretical foundation, struggle to couch their data into phenactive terms.

So far, I argued that a full turn towards radically interaction-oriented paradigms is not only problematic, but also not obligatory in order to adopt a more 'unorthodox' perspective.[5] However, there are also important insights to extract from the interactive turn. These insights mainly concern the components of embodiment, interaction and experiential quality. Although they have been brought up by traditional phenomenology, they indeed got lost when the philosophical debate focused on mindreading schemes. I agree with proponents of the interactive turn that a narrow view on the observational inference of mental states does not reflect the manifold nature of social cognition.

On the other hand, it is questionable whether phenactive theories are able to capture the whole picture or social cognition. Although they might yield ways to grasp interaction, embodiment and phenomenology, it is unclear how they would account for other elements of the phenomenon, such as offline construction. More importantly, central desiderata of a theory of social cognition are not fulfilled by radical phenactivism. In previous chapters, I have doubted the empirical plausibility of phenactive paradigms. Further, it is questionable how well a radical theory sits with interdisciplinarity. If most of the background claims of those disciplines that are central to investigating social cognition are denied, will there be room for a fruitful dialogue?

In order to methodologically implement interaction scenarios that actually reflect real-life situations of embodied agents and in order to interpret the results such methods bring forth, I claim that we do not need to commit to a radical phenactive position. If the goal is to provide a *comprehensive* theoretical framework for social understanding, it might be even harmful to adopt such a radical position (or, rather, *any* radical position). As matters stand now, it seems that both cognitivist and enactive theories have contributed valuables insights to the debate. It could be that some social processes need a rather non-representational, non-computational background, while others ask for a more cognitivist picture. I therefore argue that we should preserve a middle course and try to prevent any extreme, radical position that potentially excludes important components of the phenomenon. It would be advisable to try and find a theoretical framework that is able to integrate the full spectrum of social aspects.

[5] 'Orthodox' views, in contrast to unorthodox ones, would be cognitivist, internalist theories such as ST and TT

I thus conclude that no full turn towards enactivist theories is necessary to capture the many faces of social cognition. The interactive turn has raised awareness to neglected areas in the field. It is thus worth turning our heads to see what is on the other side. However, it might not be necessary to actually turn our backs to our already existing theories, at least not at this point. The upshot of this section is that we need a framework for social cognition that is able to coherently integrate claims from both sides of the theoretical spectrum. One way to do so is to adopt a pluralistic perspective on social cognition. In the following section, I will review already existing pluralistic accounts and argue that although the idea is highly valuable, they have a significant shortcoming. In the second part of my thesis, I suggest ways in which those can be overcome.

4.2. Pluralistic Accounts of Social Cognition

In this subchapter, I describe the very recent development in the research field on social cognition to view the phenomenon in a pluralistic manner. Proponents of such a strategy argue against the claim that there is *one* default strategy of social understanding. Instead, their basic assumption is that social cognition draws on a variety of strategies that are applied depending on situational context.

(1) In chapter 4.2.1., I present Newen's (2015) *multiplicity view* (MV) as an example of a pluralistic theory of social cognition.

(2) After describing this view in more detail, I argue that pluralistic approaches are highly valuable, because they do justice to the manifold of social understanding and offer a novel way of unification and integration. I also show that such a view is able to include demands from the interactive turn. However, there are also several caveats to MV that will be discussed.

4.2.1. The Multiplicity View

One fact about understanding others that should be clear by now is that there are many different strategies that can be applied. Which of those are chosen to solve the task at hand depends on the context individuals find themselves in. This view has been dubbed *multiplicity view* (MV) by Newen (2015) and can be seen as one step towards a reconciliation of different approaches to social cognition. In this section, I describe MV in more detail and illustrate the value of the idea for a comprehensive framework of social cognition.

As we have seen in chapter 1.4.1., the contention that there is no single all-purpose mechanism for social understanding has been introduced long before the explicit formulation of MV or other pluralistic accounts. Goldman (2006, p. 43), for example, proposes a hybrid model of mindreading, suggesting that there are “a number of ways to blend simulation and theorizing elements into a mosaic of mindreading possibilities.” Adolphs (cf. 2006a, p. 30) puts forth a similar view and argues that although simulation may constitute a good part of social cognition, it is unlikely that it is the *only* mechanism humans apply to infer the mental states of other people. Newen (2015, p. 7) takes up this line of thought and claims that

> [t]here is no standard default strategy of understanding others, but in everyday cases of understanding others we rely on a multiplicity of strategies which we vary depending on the context and on our prior experiences (and eventually also triggered by explicit training).

What leads him to this conclusion? His argument boils down to the assumption that context determines which strategy is most likely to be used. To see this, think of the following example. I enter my living room and see my friend crying and sobbing. As a neurotypical person, I immediately realize that something is wrong. This immediate hunch arises without any explicit thought or conscious inference (such as, ‘My friend is crying. Usually, people cry when there is something wrong. When I cry and sob, something bad happened. I thus infer that there is something wrong with my friend’). It is more likely that another epistemic mechanism led me to the feeling I got when I saw my friend. On the other hand, figuring out the reason for her sadness might indeed need conscious inference (such as, ‘She is saying that everything is alright, but she does not seem like it. Maybe it has something to do with her new job?’)

This example shows that the situation may already give enough context to figure out some, but not all aspects of what is going on with the other person. Newen (cf. ibid., ch. 3) claims that there are four basic epistemic mechanisms of social cognition, namely simulation, theoretical inference, direct perception, and primary interaction. He further lists constraints for each of them. As for simulation, one condition that must be fulfilled is that individuals share similarities. Theorizing, as I pointed out in my example above, only arises in rather complex social situations that require explicit thought to disambiguate the input. Simulation and theorizing, according to Newen, are high-level mechanisms which are cognitively more 'costly' than direct perception and primary interaction.[6] These latter two exhibit intuitive ways of understanding others. When encountering a person that one already has a rich amount of information about, direct perception is activated. Easily disambiguated social situations then only need primary interaction in order to be understood. It is thus concluded that "[o]nly the combination of all four strategies, in full sensitivity to the context and applied on the basis of our experience in successfully using the strategies, makes us experts in understanding others." (Newen, 2015, p. 7)

The idea just outlined has several advantages over theories that foster a default strategy for social cognition. First, it pays attention to the experiential nature of social encounters by including direct perception and primary interaction. As such, elements that have been neglected are integrated. Second, since direct perception and primary interaction are concepts that – in their original formulation – assign a fundamental role to the body (see chapter 2.1.3. and Gallagher & Hutto, 2008), this view could also be able to include the component of embodiment. Furthermore, MV factors in the role of prior knowledge individuals possess, may this be culturally acquired knowledge or information about another person that has been gathered in past interactions. MV therefore – at least potentially – fulfills important demands from the interactive turn.

However, while the general idea of MV strikes me as very fruitful and promising, there is an objection I wish to raise. I claim that Newen's strategy of putting together elements from different theoretical strands runs the risk of combining elements that are potentially incompatible.

4.2.2. Multiplicity Needs Consistency

While it certainly makes sense to assume that humans do not rely on one single strategy to navigate their social world, trying to accommodate several theoretical aspects of social cognition does not come too easy. If I am right and the interactive turn leads the research field to not adopt one radical position, but motivates to 'cherry-pick' from both sides, we will have to pay careful attention to the *compatibility* of those cherries.

[6] Newen (cf. 2015, pp. 3–4) argues against the view that simulation could be a sub-personally realized process. He builds his argument on Gallagher's (cf. 2007, p. 360) claim that simulation is a personal-level mechanism that cannot coherently be ascribed to the sub-personal level.

In previous chapters I showed that while ST and TT grew out of a cognitivist, representational theory of the mind, phenomenological and enactive theories start from the assumption that this view is fundamentally flawed. I argued that this is not only due to their historical growth, but that there are important systematic reasons for why each theory is embedded in their metaphysical framework. To briefly recapitulate the contrast between them, there are important differences in how phenactivism and cognitivist theories view the nature of the mind and the relation between the brain, body and environment.

First, consider the relation of mental processes and the external world. A rather cognitivist view assumes that causes in the world need to be inferred, the outside world needs to be *internally represented*, a job that is executed by the brain. Since other people and their mental states are part of this outside world, too, their mental states need to be inferred and represented as well. If it is taken for granted that the brain is the only mental organ (Hohwy, 2015), then it should also be clear that the *location* of the mind is to be found inside the skull. ST and TT neatly fit into this picture, since they are described as inferential processes whose function is to disambiguate social input and because both simulation and theorizing are assumed to be neurally implemented.

By contrast, a phenactive view holds very different assumptions about what and where the mind is. To see this, I once more wish to emphasize that phenactivism fosters neither an internalist nor externalist view of the mind, but a *relationalist* one. As such, it describes the mind to emerge within the interplay of agent and environment. Assuming that the mind unfolds in such a relational manner enables the view that interaction – which constitutes itself between two individuals – indeed carries the 'cognitive load' of social cognition. Such a claim is not feasible in an internalist perspective, which does not ascribe any constitutional power to mind-external features. These contradicting background assumptions are the reason why MV is potentially an inconsistent theory. It claims that social cognition involves direct perception and primary intersubjectivity, two concepts that require a phenactive view on the mind. MV also asserts that simulation and theorizing are mechanisms for social understanding, which on the other hand are based upon cognitivist views of cognition. This combination of theoretical elements is problematic, because we would have to depict the mind as relational for the one, but the other mechanism.

At this point it is useful to make explicit what this boils down to and what it means for the current work. My goal here is not to establish a metaphysical framework for the sciences of social cognition, but simply to claim that if we want to reconcile different views, we cannot merely combine aspects in a “buffet approach”- manner (“take what I like and leave the rest untouched”, Dennett, 2007, p. 248) of camps with contradictory metaphysical convictions.[7] The least that will have to be done is to pay attention to those issues and check for compatibility when needed. This can be formulated as another central desideratum for a comprehensive theory of social cognition:

d_6 – consistency
Consistency demands that a theory is built upon non-contradictory background assumptions, but rests on coherent grounds.

On the other hand, it will not be useful to ‘throw out the baby with the bathwater’ and reject either side entirely right from the start. Whether or not phenactive and non-phenactive views of social cognition can be reconciled and how we can begin to do so is an open question. It could be that there are ways to integrate them so that the research field does not have to face an either-or-question. If this is the case, however, we should ask ourselves how real the enemy at either side of the theoretical spectrum is.[8] This leaves us with two more desiderata for further research. First, we should be open to insights from both cognitivist as well as enactive views of cognition and social cognition:

d_7 – comprehension
Comprehension asks for encompassing the relevant components of social cognition and comprising elements from several accounts.

Second, a philosophical meta-theoretical framework can help to distill possibilities for integrating those insights which are deemed most useful:

[7] For an extensive defense of the importance of metaphysics for (cognitive and behavioral) sciences, see Ross, 2004, p. 606: “Our talk about “scientifically interesting metaphysics” gestures at the following fact. It *is* a feature of scientific epistemology, as really practiced in laboratories and journals, that the various pieces of scientific inquiry must broadly cohere into a general world-view that, at least in its core, almost all signed-up members of the mainstream scientific professions can share. Furthermore, it is a legitimate job of the “serious” metaphysician to ensure that proposals for articulating and enriching this world-view are, at least potentially, genuinely enlightening, and not merely verbal or technical.”

[8] See Dennett (2011) for a review of Thompson’s (2010) book *Mind in Life.* Dennett claims that enactivism is, after all, not that controversial and merely rephrases claims that most cognitivist agree on.

d_8 – integration
Integration postulates that a philosophical theory of social cognition must be able to include elements from a variety of interdisciplinary accounts on the phenomenon.

One of the biggest advantages of pluralistic theories, it seems to me, is that they are in principle able to capture ideas from both ends of the theoretical spectrum and thus to contribute a valuable framework in our educational journey. Integration thus is a crucial desideratum for a theory of social cognition, and one that should fall into the realm of philosophical work.

This leaves the interdisciplinary research field with the task of reconciling and making conceptually consistent our choice of a specific, unified methodological framework. In other words, our overarching theoretical approach to components of social cognition that have been described by multiple disciplines needs to be formulated. In doing so, it should be a joint objective to operate on a consistent set of metaphysical assumptions, since this choice has forming consequences and implications for both theoretical and empirical research.

Part 2

5. Building Block I: First-, Second-, and Third-order Embodiment (1-3E)

> *What is a "grounding relation"? A grounding relation connects a given intentional and/or phenomenal target property with the situated, low-level physical dynamics of a given type of cognitive and/or conscious system, for example by specifying a computational model that allows us to understand the transition from the representational or phenomenological level of description to the one of physical microdynamics. (Metzinger, 2014a, p. 275)*

In this chapter, I illustrate Metzinger's (2006, 2014a) theory of first-, second-, and third-order embodiment (1-3E) and modify it so to alleviate its shortcomings and make it more suitable for the application of social cognition. 1-3E will be the first of two theoretical building blocks that serves as the basis for my own proposal. As a hierarchical framework, it is especially useful for my purposes, providing the means for a multi-level analysis in order to integrate low- and high-level mechanisms of social cognition.

The framework has originally been developed for exploring how phenomenal properties are physically and computationally grounded. In general, 1-3E aims to yield a scaffold for analyzing and potentially explaining a phenomenal target property and to 'track down' its representational and physical groundings. Very generally speaking, Metzinger introduces three conceptual tools that are supposed to enable a more detailed analysis of the relations between different levels of description.

I will first explain the goals of this framework and describe it in more detail (chapter 5.1.). In a next step, I attempt to critically assess some aspects of it and highlight those parts that will be most important for my own suggestions in section 5.2.

5.1. Theoretical Basics of 1-3E

In this subchapter, the goal is to embed the framework in a broader context and detail its basic assumptions and goals.

(1) In section 5.1.1., I depict how 1-3E can be seen as part of Metzinger's self-model theory and the specific problem it targets. I describe two ways to use 1-3E, namely as the 'systems' and the 'hierarchical' view.

(2) Chapter 5.1.2. entails a description of the three levels of embodiment and their possible relations.

(3) One point of the theory will be especially important for my own proposal and is thus depicted in greater length in chapter 5.1.3, viz. the concepts of transparency and opacity. I explain their meaning and conceptual value for the framework.

5.1.1. The Theoretical Scope

In his earlier work, Metzinger (2004) claims that the self is not a substance, but rather the phenomenal end-product of a complex representational process which happens to take place in embodied systems. This assumption raises the question of how exactly the experience of being a self is generated by processes of embodiment; in other words, what are the grounding relations of phenomenal selfhood?

> It is the problem of describing the abstract computational principles as well as the implementational mechanisms by which a system's phenomenal self-model (PSM; Metzinger, 2003; 2007 is anchored in low-level physical dynamics, in a maximally parsimonious way, and without assuming a single, central module for global self-representation. (Metzinger, 2014a, p. 272)

Metzinger's goal for 1-3E is to show how the experience of being a self (i.e., phenomenal selfhood) is generated within an embodied system (cf. ibid.) and thus to embed his self-model theory (Metzinger, 2004, 2007) in the context of grounded cognition.

The basic claim of his approach is that phenomenal selfhood is grounded in computational, representational processes that in turn are grounded in physical (i.e., neural and bodily) structures. This threefold structure is reflected in three orders of embodiment that in general refer to different levels of description; experiential phenomena such as phenomenal selfhood have a specific phenomenal quality, which can be described at the level of third-order embodiment (3E), the phenomenological level of description. They are functionally realized by processes that are described as representational or computational. This is referred to as second-order embodiment (2E), the computational or representational level of description. The implementational level of description then exhibits the physical grounds of those processes at the level of first-order embodiment (1E). By tracking down the computational and physical counterparts of phenomenal states, Metzinger aims to draw a picture of the grounding relations between them. To emphasize that the body plays a crucial role for phenomenal selfhood, the different levels of description are re-formulated as levels of embodiment; namely first-, second-, and third-order embodiment.

Importantly, there are two readings of 1-3E, which I will call the ‘systems’ reading (Fig. 2) and the ‘hierarchical’ reading (Fig. 1). The latter reading exhibits orders of embodiment *within one system*.

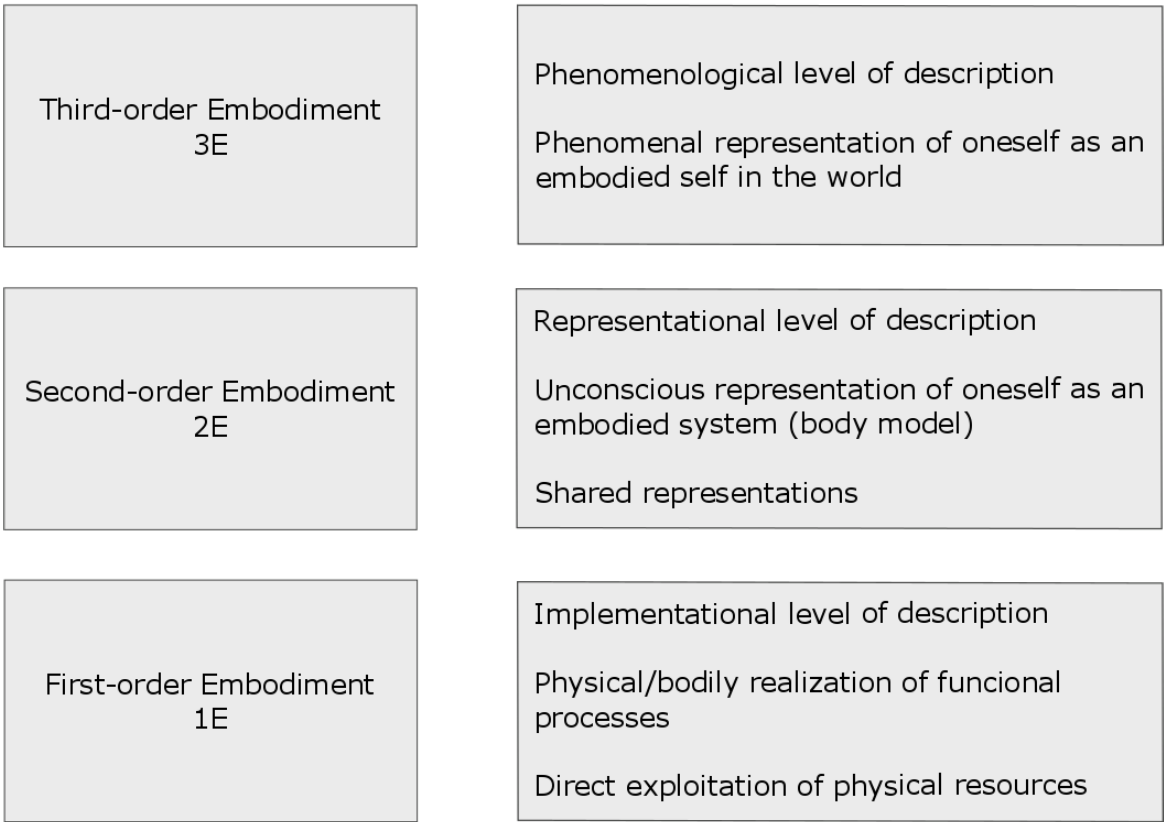
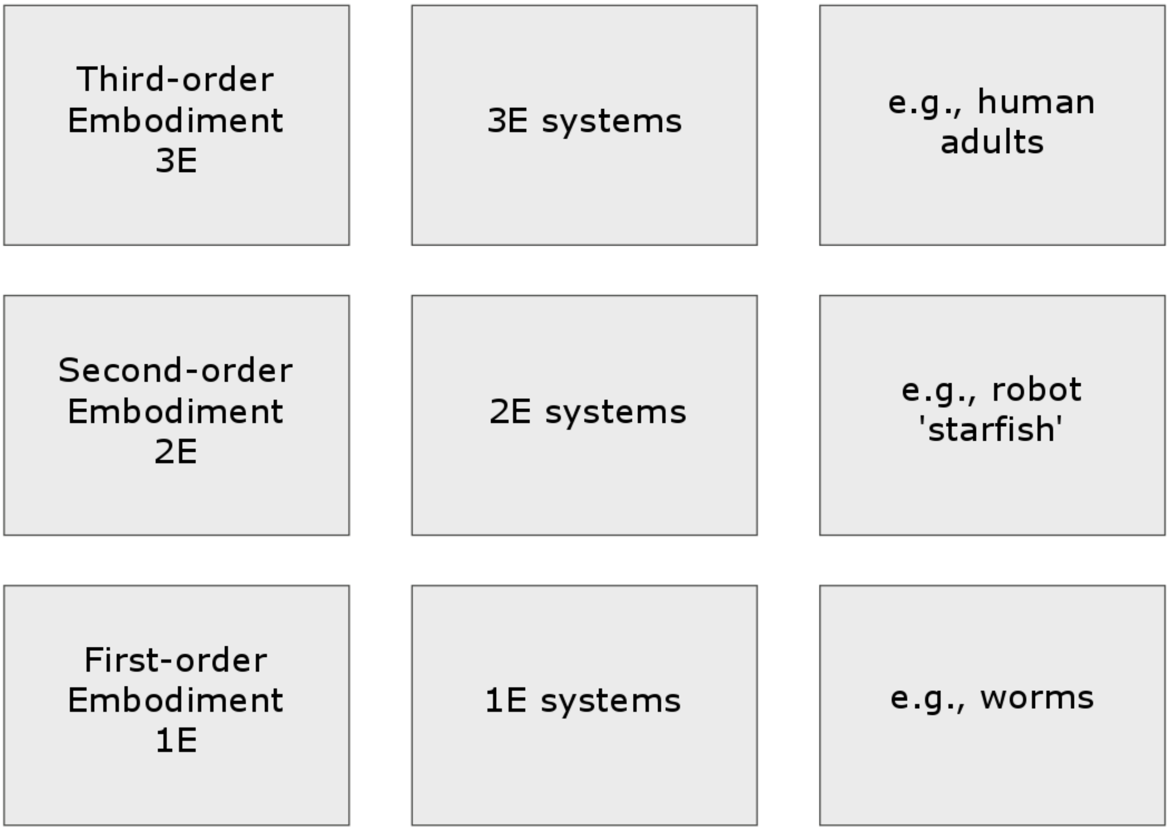

Third-order Embodiment
3E
Phenomenological level of description
Phenomenal representation of oneself as an embodied self in the world
Second-order Embodiment
2E
Representational level of description
Unconscious representation of oneself as an embodied system (body model)
Shared representations
First-order Embodiment
1E
Implementational level of description
Physical/bodily realization of funcional processes
Direct exploitation of physical resources

Fig. 1 The hierarchical reading

The hierarchical reading exhibits levels of embodiment as processing stages within one system and as levels of description. It is thought that the levels build upon each other, in the sense that every 2E system possesses 1E, and every 3E system possesses 1E and 2E.

The two readings relate in that it is assumed that each 2E and 3E system possesses lower levels of embodiment as grounding relations. To see this, consider three examples: a worm, an advanced robot (e.g., the “starfish”, see Metzinger, 2007), and an adult, neurotypical human. They all possess a body and some sort of more or less sophisticated, skillful behavior which allows them to interact with their environment. The different levels of embodiment can now be interpreted as different levels of sophistication, both within one system as well as in a ‘hierarchy’ of systems.

After this brief introduction, I will now describe the theory and different orders of embodiment in more detail.



Third-order Embodiment 3E	3E systems	e.g., human adults
Second-order Embodiment 2E	2E systems	e.g., robot 'starfish'
First-order Embodiment 1E	1E systems	e.g., worms

Fig. 2 The systems reading

Orders of embodiment individuate classes of systems. A system that possesses 3E, for example a human in a non-pathological waking state, is described as a 3E system. A 2E system possesses second-order embodiment. An example is the robot ‘starfish’, which unconsciously generates self-models for motor control. 1E systems, such as worms or simpler robots, possess first-order embodiment

5.1.2. First-, Second-, and Third-Order Embodiment

Let me first spell out the ‘systems’ view by referring to the examples mentioned above. Considering the worm, it can be assumed that it directly exploits its physical resources to navigate its environment. However, it is rather unlikely that in doing so, the worm makes use of any rule-based computation over explicit symbol-like representational structure that could be found in its nervous system. Systems like the worm, which do not exhibit any (explicit) representational or phenomenal states, are described as ‘1E systems’, since they possess first-order embodiment. According to Metzinger (cf. 2014a, pp. 272–273), 1E systems can be conceptualized within dynamical systems theory (Chemero, 2009) or microfunctionalism (Clark, 1990) and as such can be described as systems whose mental properties emerge from the direct interaction of body and environment. In contrast to this rather rudimentary kind of embodied agent, systems that possess second-order embodiment ('2E systems') are representational systems that unconsciously represent themselves *as* embodied. This unconscious representation is described by Metzinger (cf. 2014a, p. 273) as a unified, global representation of one's body, the body model. This body model can be exploited in several ways, for example as a functional tool for predictive motor control, and sustains interactions with the environment. 2E also enables counterfactual representation, that is, the capacity to represent potential states of the system without their actual execution. We will see that this ability is crucial for a range of social cognitive processes. The body model thus can be said to functionally underlie both physical and virtual behavior (see Cruse & Schilling, 2015). Importantly, the body model also forms the basis for shared representations with other agents. I will detail this claim further in chapter 2.1.2.

What 2E systems lack, however, is the ability to *phenomenally* represent themselves as embodied agents. While a robot like the starfish can exploit and use its unconscious body model for movement and other skillful behavior, it does not experience itself as doing so. This phenomenal quality only arises in 3E systems; the distinctive feature of that kind of system is that they *experience* themselves as embodied agents that own and control a body. Humans in non-pathological waking states are one example of 3E; along with their ability to exploit their body model comes the phenomenal experience of having control and *owning* their body (cf. ibid., pp. 274–275).

The just depicted systems view relates to the hierarchical view in that it is assumed that each 3E and 2E system possesses the respective prior orders of embodiment. In this way, it can be said that 1-3E is a grounding theory: the phenomenal properties that 3E systems have are computationally grounded in a unified representation of the body (2E). This unconscious body model, in turn, is grounded in physical, bodily resources that are described at the lowest level of the hierarchy, 1E.[1] The theory thus allows to take one target system, or a target phenomenon of one system and describe what different levels of embodiment within this kind of system amount to and how they relate. Phenomenal selfhood in human adults, for

[1] For a discussion of the possibility that 3E exists independently of computational and physical grounding relations, see Metzinger, 2014, p. 278.

example, would be a target that is localized at 3E. How is it brought forth by representational processes, and what kind of computation does such a system need at the level of 2E? What are physical structures that subserve those processes, or maybe even enable or constitute them? Which structures and processes are necessary, which are sufficient? Those are the kinds of questions for whose answers 1-3E aims to yield the conceptual tools.

5.1.3. Transparency and Opacity

There is one important aspect of 3E that shall now be described in more detail, since it will be crucial for my own account. As I have shown before, the feature that sets 3E and 2E systems apart is that only the former possess phenomenal properties that are brought forth by computational processes. According to Metzinger (2003), there are two kinds of phenomenal properties that can be instantiated by conscious representational states; transparency and opacity. Let me briefly detail the usage of the terms in this context.[2]

As a helpful analogy, think of sitting at your desk and looking out of your window. First, imagine you just cleaned this window the day before, so that it is nice and clean. In this case, the glass is *transparent*, that is, you look right through it onto your garden or whatever is behind that window. That is to say, you do not experience the window *as a medium* you are looking through. Now imagine that cleaning day still lies ahead and your window is pretty dirty. When looking out, you most probably not just see what is behind the window, you are also aware of the fact that you are looking through glass – the window is opaque and due to this property, you perceive it as a medium that allows you to see your garden.[3]

In analogy, Metzinger claims that mental states and their processing stages are either transparent or opaque. A mental state is opaque when an individual experiences it *as a representational state*. Conscious thought is a pretty straightforward example; I am now aware of the fact *that I am thinking*. The representational process itself is represented – just like the dirty window you experience your own medium of thought, hence the attribute 'opaque'. Along with this experience comes the awareness of the possibility of

[2] Metzinger uses the terms as originally introduced by Moore (cf. 1903, p. 446). A more detailed description of alternative usages of the terms transparency and opacity, see Metzinger, 2003, pp. 354–358.

[3] Metzinger (2003, p. 358) makes use of a similar example: "With regard to the phenomenology of visual experience transparency means that we are not able to see something, because it is transparent. We don't see the window, but only the bird flying by."

misrepresentation – the awareness that my representation could be wrong. If a mental state is transparent, in contrast, the awareness that this is a representational process is not part of the phenomenal content: "Phenomenal transparency in general, however, means that something particular is not accessible to subjective experience, namely, the representational character of the contents of conscious experience." (Metzinger, 2004, p. 169) Rather, one experiences oneself as immediately and directly being immersed in a world – without any awareness that there is something like a medium which enables this impression.

What this distinction offers, it seems to me, is the possibility to yield a fine-grained description of different types of phenomenological experiences. It thus takes the phenomenological level of description seriously and relates it to 'lower' or prior stages of processing. Importantly, transparency and opacity are no all-or-nothing phenomena. Much rather, phenomenal states can oscillate between being transparent and opaque, or may be transparent and only partially opaque. I discuss this possibility in greater lengths in section 7.4.2. In chapter 7.4., I further apply these terms to the phenomenology of social encounters and by doing so hope to give a more detailed perspective on different kinds of social mental states.

5.2. Discussion

In this section, I wish to discuss, clarify and modify several aspects of the theoretical ideas described above and show possible ways to enrich Metzinger's theory. Accordingly, these modified ideas will be applied to the study of social cognition and described in more detail in the following chapter. The present section will include the discussion of the following aspects:

(1) In chapter 5.2.1., the meaning of embodiment in the framework of 1-3E and the depiction of relations between different levels of embodiment are scrutinized and critically assessed.

(2) The possibility of combining representational and non-representational levels of description is discussed in chapter 5.2.2, pointing to a possible solution to the problems may that arise.

(3) The last subchapter introduces an additional level of embodiment (3E+) to strengthen the distinction between transparent and opaque phenomenal states. I present why such a stronger distinction improves the descriptive power of the framework.

5.2.1. The Notion of Embodiment

The title of the framework suggests that embodiment plays a fundamental role for cognitive processing. Metzinger emphasizes that there are grounding relations holding between first-, second-, and third-order embodiment. In order to shed light on those relations, it will be helpful to unpack the jargon first. Metzinger (2014a, p. 277) borrows Pezzulo and colleague's (2013) conception of a grounding theory of cognition. Here, it is claimed that cognition is grounded, embodied, and situated. It is grounded because it "has a physical foundation" (ibid., p. 4), it is embodied because bodily processes profoundly shape cognition. These bodily, i.e., sensory and motor, processes are "structured according to physical principles that provide the grounding of cognition" (ibid.). Situatedness of cognition, in this sketch, refers to the fact that cognition occurs in given contexts that shape and influence the process at hand. Taken together, the authors assume the following: "For this, we propose that the effects of grounding, embodiment, and situatedness can be conceptualized as a cascade and have additive effects on cognition and representation." (ibid.)

Metzinger (cf. 2014a, p. 277) scrutinizes this claim and explains that there are three steps to take in arriving at a depiction of grounding relations for phenomenal properties. First, phenomenal properties have to be described in terms of their representational properties. Those have to be 'bottomed out' by formulating a computational model in a second step. The third step then is to find the necessary, enabling and constitutive parts of a given phenomenon. This profoundly helps to extract the actual grounding relations, for "[g]rounding is about constitution" (ibid., p. 277). He further claims that 2E and 3E (in our world) are grounded (they "ride" on, ibid.) in 1E. However, it is important to notice that the author does not aspire to give a full-fledged, metaphysical analysis of 1E, 2E and 3E. Rather, he suggests that at this point, they should be considered as conceptual tools, that is, epistemic concepts which enable researchers to ask more detailed and specific questions about the necessary and sufficient grounding relations between them (cf. ibid., p. 278).

Let me now go into more detail about his suggestions and their potential shortcomings. While Metzinger (ibid., p. 274) is quite clear about the relation between 2E and 3E – the representational *content* of 2E is "elevated to the level of global availability and integrated with a single spatial situation model plus a virtual window of presence" – it is less obvious how 1E relates to 2E and 3E. In the general description of his theory, he does not yield a very detailed picture of what *actually* grounds 2E and which role the body plays for computation. Given that the labels he uses to coin his concepts suggest a fundamental and central role of embodiment, it seems that he does not pay quite enough attention to this topic.[4] The reason for this shortcoming may be his focus of attention, as he clarifies in a reply to Gallagher (cf. Metzinger, 2006, p. 2); his main interest lies in the relation of 2E and 3E. While this seems to be a fair enough reason, something profound, it seems to me, does not get explicitly spelled out. Let me explain this in more detail.

A 1E system is described as a "purely physical, reactive system" (Metzinger, 2014a, p. 273) that adapts to its environment by exploiting its physical resources. This is not, however, what is represented at the level of 2E, that is, what grounds the unified body model. It is merely a description of a 1E *system*, but not a description of 1E as the grounding level *within* one system. To see this, remember that Metzinger offers a description of the relation between 2E and 3E; some of the representational content is now available for phenomenal experience. It seems, though, that the theory lacks such a description for 1E and 2E. What is it, really, that grounds the computations?

In an earlier paper, Metzinger (cf. 2006, p. 3) suggests that many human behaviors are directly shaped by our bodily configuration, such as the degree of elasticity of our muscles. This gives us an idea what the lowest level of description (1E) in the framework may amount to. The physical body itself, according to this line of thinking, determines higher-level processes in that it constrains and forms the kind of cognition an individual needs. This fact is not to be underestimated. For it is rightfully emphasized by Madary (cf. 2015, p. 4), that an agent strives to gather information about the world that relates to its own sensorimotor trajectory. This means that a system's phenotype determines which piece of information is salient and which is not. A common example is the comparison of a fish and a human. For a fish, it is very important to be in water, and not to be on land to sustain its existence. However, the contrary is true for humans. It thus makes sense to say that the biological needs of an organism determines the 'kind of mind' it needs. Put in simple terms, fish need minds that keep them away from land, humans need minds that keep them on land.

[4] However, Metzinger (cf. 2014, p. 276) exemplifies a more detailed suggestion on the role of the body by reference to phenomenal dream states. Physical eye movements, in this instant, are held to ground the phenomenal experience of lucid dreams.

What is thus the relation between 1E and 2E? I suggest that 1E relates to 2E in that the former constrains the latter profoundly. 1E is the level of description where we find an agent's physical structure that determines what is and can be represented at 2E. Those structural elements do not have to be represented at 2E, rather, they constrain and influence the structure of representations at 2E. If we assume that 2E amounts to a virtual, unconscious body model that contains an integration of multimodal inputs, then 1E sets up its basic structure.

I claim that this aspect of the framework needs clarification, at least for my more specific topic on social cognition. In chapter 7, I will show how the application of predictive processing (PP) can help with this issue.

5.2.2. Combining Non-Representational and Representational Views

One caveat to the position I am about to take by adopting 1-3E and applying it to social cognition is the following. I have argued in Part I (see chapter 4) that a simple combination of representational and non-representational perspectives results in a metaphysically incoherent account. Why should it now be possible to combine 1E – a level at which it is claimed that no explicit representational processing is found – with 2E and 3E? The latter two heavily rely on representations and as such may need a different kind of approach to yield a coherent description. There are several ways to answer this valid question.

First, it is possible to argue that these are merely *methodological, epistemic* claims and as such do not imply any *metaphysical* assumptions. Framing the different levels within different theoretical approaches could yield us the best way to explain their specific properties, and that is all that matters if we want to get the best way to study the phenomena in question. This approach makes a lot of sense, it seems to me, for the systems view of 1-3E (see chapter 5.1.1.). Methodologically, it is possible to investigate 1E systems in the light of phenactive or dynamical system theories, which focus on a heavy reliance of the relation between mind and world. Systems that are more complex, on the other hand, may be more fruitfully accounted for by a theory that assumes that there is some type of computation or representation that underlies this complexity.

However, it gets trickier, in my opinion, when trying to apply this strategy to the hierarchical view of 1-3E. Especially as a philosopher I opt for paying a minimum of attention to theoretical and metaphysical consistence. I find it highly problematic to apply a phenactive theory to lower processes, while using theoretical claims of a cognitivist theory for the higher ones. The problem I see is that we would have to assume two different kinds of minds for different processing stages *within one* system (with, most probably, one mind). To assume that the mind is relational and non-representational for some processes, and then to assume that the mind is fully internal for other kinds of processes appears contradictory, or at least like a very unparsimonious and cumbersome description.

Metzinger (2014a, p. 278) suggests a way to solve this problem and claims that representations should be seen as a gradually emerging phenomenon:

> As indicated above, my own recommendation would be not to throw the baby out with the bathwater, but to mathematically describe "representationality" as a property that is graded, allowing the PSM [phenomenal self-model] to naturally "bottom out" into non-representational dynamics.

Another way to solve this problem is to claim that different processing stages recruit different structures – some more crude and physical, others more sophisticated and abstract. Given that both human and non-human animals exploit not only their bodies, but also artefacts in their environment to solve problems (e.g., Menary, 2012), it is plausible to assume that cognitive systems are "leaky" systems:

> […] "leaky" in the sense that many crucial features and properties depend precisely on the interactions between events and processes occurring at different levels of organization and on different time scales. (Clark, 2014, p. 248)

With this kind of view, the basic cognitive machinery is located within the agent, but still allows for a tight relation and dependence upon brain-external devices. In my own proposal, I will pick up these ideas and argue in more detail for them. I claim that PP will be of great help to depict such a gradual view of representations and to assign a role to non-neural structures. Although being a strongly representational framework, PP can be interpreted in a way that opens the possibility of a tight relation to non-representational, real-world and bodily structures (e.g., Clark, 2015b, 2015c; Seth, 2015b).

5.2.3. Transparency and Opacity Revisited

My last comment on the original framework concerns the distinction of transparent and opaque phenomenal states. While I want to keep the terms of transparency and opacity and think that they are helpful tools for gaining a more detailed description of phenomenal properties, there is one point that shall be revised. I opt for a stronger conceptual reflection of the distinction between transparent and opaque states, since there seems to be an important qualitative difference between them.

In general, transparency can be viewed as the ‘ground level’ of consciousness, while opacity appears to be a much more rare and specific kind of experience. Transparency enables phenomenal representation and self-identification. It allows a system to identify itself with the model that is generated, particularly in virtue of the fact that the system cannot detect this modelling process. It is ‘tricked’ into experiencing itself as being in direct contact with the world, without experiencing any sub-personal generation or construction process. It is easy to imagine that non-human animals or human infants, unable to form explicit beliefs about the world, have such phenomenal properties. However, the insight that I am a thinking and representing system, i.e., the property to detect representations *as* representations, seems to be restricted to a small subgroup of the human species. Even human infants or human adults with specific psychiatric or neurophysiological impairments may not have this kind of experience.

Metzinger (cf. 2014a, pp. 273–274) argues that the distinctive feature of 3E is the possibility of a system to identify itself with its body, resulting in the phenomenal properties of selfhood and self-identification. This speaks for an interpretation of transparency and opacity not as mutually exclusive properties, but much rather implies that transparency is a property of *any* phenomenal state. If that is the case, the following consequences ensue:

On the one hand, it appears that opacity is an *additional* kind of representation (i.e., added to the transparent process). At the same time, the representation process is most likely never *entirely* represented, thus always leaving some of the process transparent. How are those two consequences to be integrated? My proposal is twofold. First, I introduce the level of 3E+, which describes opaque phenomenal states. 3E+ is supposed to conceptually grasp that opaque states are qualitatively different from transparent states (3E). Secondly, it needs to be considered that characterizing opacity as an additional kind of representation does not mean that each opaque state is a *meta-cognitive* state.[5]

To see this, consider the following example: You are sitting on your couch, completely immersed in a daydream. Your phenomenal state is fully transparent, there is no awareness

[5] Meta-cognition is here used as “thinking about thinking”, “cognition about cognition”, etc. This is not to be confused with meta-representation. According to Metzinger (2004, p. 33), any kind of introspection is meta-representation, since it is “the internal representation of active mental representata.” For a more detailed description of different kinds of introspection, see ibid., p. 36.

of the fact that you are caught in a daydream. However, at some point your attention shifts towards that fact, you become consciously aware that you were daydreaming.

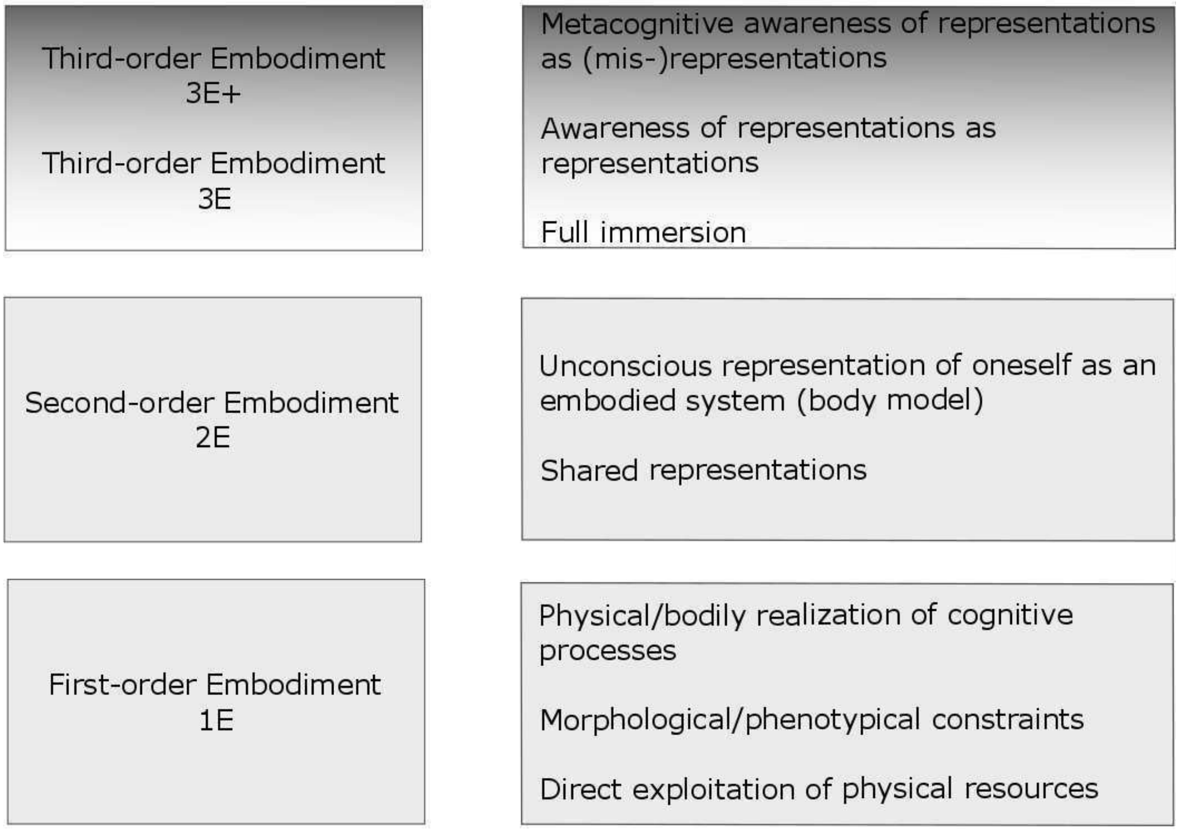

Third-order Embodiment
3E+
Third-order Embodiment
3E
Metacognitive awareness of representations as (mis-)representations
Awareness of representations as representations
Full immersion
Second-order Embodiment
2E
Unconscious representation of oneself as an embodied system (body model)
Shared representations
First-order Embodiment
1E
Physical/bodily realization of cognitive processes
Morphological/phenotypical constraints
Direct exploitation of physical resources

Fig. 3 Revised hierarchical reading

At the level of 3E, an additional level has been added in order to emphasize the difference between transparent and opaque phenomenal states. 3sE+ describes states during which a metacognitive awareness of representations as (mis-)representations is present.

This is when "[…] we consciously represent that something actually is a representation, not by propositional knowledge or a conscious thought, but first by our attention being caught by the fact *that* what is currently known is known through an internal *medium*." (Metzinger, 2004, p. 170) While this can already be described as an opaque phenomenal state, a 'stronger' version of opacity is at play when you now deliberately decide to stop daydreaming and steer your thoughts towards something else. It thus seems to me most useful to describe 3E and 3E+ as a spectrum of phenomenal qualities, ranging from fully transparent to strongly opaque states. The more explicit the construction process is represented, the closer a systems moves towards to 3E+ end of the spectrum.

This revised model (Fig. 3) enables a more fine-grained description of phenomenal experiences and emphasizes the uniqueness of meta-representation.

In what follows, I aim to apply this modified version of 1-3E to the study of social cognition and show how the consultation of PP offers new and exciting ways to overcome possible issues.

6. Building Block II: Predictive Processing (PP)

> *The brain thus revealed is a restless, pro-active organ locked in dense, continuous exchange with body and world. Thus equipped we encounter, through the play of self-predicted sensory stimulation, a world of meaning, structure, and opportunity: a world parsed for action, pregnant with future, patterned by the past. (Clark, 2016, p. 300)*

Predictive processing (PP) is a growingly prominent and promising theory in the cognitive sciences that aims to provide a unifying principle of how perception, action and cognition are brought forth. It is the second theoretical building block for my framework that I will develop in chapter 7. The reasons to choose PP as part of my own theory are manifold. First, it sits well with 1-3E, but offers important additional features. Additionally, it provides an interesting twist of the notion of representations, giving them a closer relation to embodiment and action. This is one general advantage of the version of PP I foster here, it is able to smoothly integrate theoretical insights from both cognitivist as well as phenactivist views. In doing so, however, it stays on metaphysically coherent grounds and thus is the perfect partner for my aspirations. Before going into more detail on the many ways in which PP can be exploited for social cognition (I will do so in chapter 7), let me first describe the theory itself.

The rationale of PP is that there are hidden causes in the world – hidden for the brain in the skull – that need to be inferred (chapter 6.1.). All the brain can access are its own states, it can neither directly access the external world, nor the body it is embedded in, nor other agents. It thus has to solve an *inverse problem*, viz. how do the effects on the brain relate to causes in the external world (cf. Hohwy, 2013, p. 53). This perspective seems *prima facie* rather brain-bound and internalist, and indeed it has been claimed that adopting PP results in returning to strict internalism (Hohwy, 2014a). However, as I wish to show throughout this chapter, there are also convincing positions that relate PP to embodiment and the continuity of mind and life. As an instance of the free energy principle (FEP, Friston, 2010), PP naturally connects life and mind, one core demand of phenactive views, as will be described in chapter 6.2. To put it in Seth's (2015b, p. 5) words:

> This view of the brain is shamelessly model-based and representational (though with a finessed notion of representation), yet it also deeply embeds the close coupling of perception and action and, as we will see, the importance of the body in the mediation of this interaction.

PP thus appears as an extremely useful theory when it comes to putting key insights of phenactive and cognitivist theories together. The aim of this chapter is to introduce PP in more detail and to investigate the ways in which I think PP is indeed the unifying perspective it has been praised to be (e.g., Clark, 2013; Friston, 2010; Hohwy, 2013).

Although the brain plays the main role in cognitive processing, embodiment and action are just as fundamental for PP (chapter 6.3.). They are an indispensable part of prediction error minimization, which lies right at the heart of the account. PP has been convincingly extended to interoceptive processing, and thus includes important insights about the role of emotions (Seth, Suzuki & Critchley, 2011). This will be scrutinized in chapter 6.4. Finally, while the account fosters a deep connection of brain, body, and environment, 'representation-hungry' phenomena such as counterfactual thinking, imagination and dreaming can be explained quite naturally. In chapter 6.5. I discuss the advantages of PP in more detail and come to the conclusion that it not only provides the means to supplement 1-3E, but that it also sheds new light on the vivid debate on representations and the mind in more general.

In what follows, I will flesh out the specifics of this theory and hope to show that PP is indeed a promising candidate for framing social cognition in virtue of its unifying power. Before I describe the claims that help us "putting predictive processing, body and world together again" (Clark, 2015a, p. 11), let me briefly summarize the most important assumptions of the theory.

6.1. The Basics

In this brief chapter, I will introduce the basics of PP and circumscribe its problem space:

(1) In chapter 6.1.1., the basic problem that biological agents are faced with when it comes to figuring out causes in the external world is described. I will detail the idea that through picking up regularities in the world, the most probable cause of an incoming sensory signal is inferred.

(2) Chapter 6.1.2. depicts the hierarchical architecture of PP and shows that the most important task of the brain is to minimize prediction error. The role of probability and prior knowledge to find the most likely cause of an effect are scrutinized.

6.1.1. The Brain’s Task

As already briefly mentioned, PP assumes the following fundamental problem for organisms. An organism’s brain is confronted with a multitude of sensory signals, each of which has multiple possible causes. To see this, consider the following scenario: You wake up in the middle of the night, your room is pitch-dark. Something soft touches your forehead. Is it your cat, or your partner’s hair? Your pillow or blanket? A hairy spider? In this example, the sensory input (something soft on your forehead) can have multiple causes. The brain itself is locked inside the skull, it cannot access the external world and the hidden causes out there directly. This means that the brain needs to *infer* what is out there. At this point, let me clarify the notion of inference that is used here. PP is mostly a theory about the way the brain computes at a sub-personal level. It offers insight about how the system comes to find the most likely cause of an incoming signal. The example of the fluffy sensation on your skin is not supposed to suggest that the kind of inference that is needed to find out what touched you is made consciously and at a personal level. To the contrary, most inference that will be talked about in relation to PP will be sub-personal. In this sense, Brown and Brüne (2012, p. 2) state that “[i]nference can refer to deterministic short-term processes that are largely situated in current behavior, and are probabilistic estimations about the state of the world, and are most relevant to prediction errors and concepts models with Bayesian statistics (Friston et al., 2009).” I will go into more detail about how the sub-personal and personal level relate and which novel insights about this relationship are given by PP. For now it shall be sufficient to note that inference in this scheme is more often than not an unconscious, computational process.

The world we live in is full of causal regularities that come at different spatial and temporal timescales (cf. Hohwy, 2013, pp. 27–28). Examples of such regularities are ‘what goes up must come down’, or ‘day comes before night’. However, there is also noise and irregularities, unpredictable, surprising events or disturbances of signals. The brain’s task is now to find the regularities, that is, “the way causes interact and nest with each other across spatiotemporal scales.” (ibid., p. 28) Knowing about causal regularities is intuitively adaptive – my chances to survive are much higher if I am able to predict and cope with the environment surrounding me.

PP offers a compelling account on how a system filters out these regularities. It generates predictions about which input the world is most likely to throw at it next. These predictions are then compared against the actual signal, and in case of a mismatch, an error signal (prediction error) is formed. This error signal then is used to improve the model, so to enhance its grip on the world. In this way, actual external signals shape and fundamentally alter predictive models. It is this process that allows the brain to 'fold in' the causal regularities of the external world into its predictions. Here we already get an idea of how PP depicts a tight relationship between internal processing and extra-neural events.

6.1.2. Hierarchical Predictive Processing

Let us look at the principle of PP in more detail. First, note that it is assumed that there is a functional and neural hierarchy in the brain. This cortical hierarchy implements generative models, where each level predicts the sensory signals at the level below. While low levels predict basic sensory properties of incoming signals at fast timescales, higher levels predict more complex regularities at slower timescales (cf. Hohwy, 2010, p. 136). The basic idea now is that perception is steered fundamentally by top-down processing and that sensory information is only contained in error signals that occur when there is a mismatch between top-down predictions and bottom-up sensory input (cf. Clark, 2015c, p. 3). What exactly does that mean?

In the process of generating hierarchical generative models, hypotheses are passed down to lower levels, where they are compared to the actual sensory input. If there is a mismatch between prediction and input, a prediction error occurs. This error signal contains only the discrepancy, and it is this crucial information that now is forwarded to higher levels. Trying to improve the next prediction according to the error signal, the original prediction is adjusted and passed down again. This process goes on until prediction error is small enough, i.e., at its expected size. Remember that there are not only causal regularities in this world, but also noise and irregularities. The system thus factors in how 'reliable' a given sensory input may be, given its expected level of noise. Prediction errors are minimized until they reach their expected level of precision.

This process of prediction error minimization (PEM) lies at the heart of PP – it is the primary task of the brain. For the more errors occur, the less accurate the models, and the less accurate the estimate of the causal structures in the world will be.[1] There are several points in this

[1] While it has been argued that the ideas that I have briefly summarized above are able to explain phenomena such as binocular rivalry (Hohwy, Roepstorff & Friston, 2008), schizophrenia (Fletcher

brief description that need some more elaboration. First, consider that PP is a Bayesian theory[2] and as such highly draws on probability.

Hohwy (cf. 2013, pp. 16–17) neatly spells out three central notions: the likelihood of an event to occur, the prior probability of this event, and the posterior probability of a hypothesis. Likelihood refers to the probability that some sensory input is caused by a specific event (e.g., I wake up and feel something fluffy, warm on my face. Is it more likely that this is caused by my cat who likes to sleep on my face or by my partner who forgot to shave the past weeks?) Since sensory inputs can have many causes that all are very *likely*, the *prior probabilit*y of the event is factored in additionally.

Generative or predictive models are formed on the basis of these empirical priors, that is, prior knowledge about the world. This kind of knowledge is called 'empirical' because it is no static, inflexible kind of information, but goes through a constant updating process itself. This process is described by Clark (2014, p. 231):

> For within a hierarchical scheme the required priors can themselves be estimated from the data, using the estimates at one level to provide the priors (the background beliefs) for the level below. In predictive processing architectures, the presence of multilevel structure induces such "empirical priors" in the form of the constraints that one level in the hierarchy places on the level below.

These constraints can be, for example, shaped by sensory input. Additionally, priors are subject to a number of top-down influences. This prior probability is the "probability of the hypothesis prior to any consideration of its fit with the evidence" (Hohwy, 2013, pp. 16–17,

& Frith, 2009), and the RHI (Hohwy & Paton, 2010), there is one important, yet still unanswered question: What is the evidence for PP? As is stated even by the most convinced proponents of the theory, evidence is growing, but still indirect. It is assumed, however, that there are two units, or processing elements at each level of the hierarchy that are incorporated by two distinct populations of neurons: representation units and error units: "[…] the representation units are driven by error units at their own level and at the level below, and they send predictions both laterally and downward (i.e., to the layer below). The error or "surprise" units talk to representation units at their own level and the level above." (Clark, 2014, p. 233) These units have been hypothesized to be realized by different kinds of neurons. While error units are realized by superficial pyramidal cells, representation units are 'carried' by deep pyramidal neurons. Those pyramidal cells pass information between cortical areas; superficial neurons are responsible for forward messaging, deep cells for passing information downward (cf. Mumford, 1992, p. 243). Pyramidal cells thus have been claimed to be ideal candidates for the implementation of PP.

[2] "Bayesian statistical inference is a mathematical method of inference which incorporates priors, or prior beliefs learned from previous experiences that generate internal models of a predicted outcome, and consequently act as top-down modulators of bottom-up sensory input." (Brown & Brüne, 2012, p. 3)

e.g., the fact that my cat is lying on my face is more probable, because this has happened more frequently than my partner forgetting to shave his face).

To sum this up, there are two ways to find out a hidden cause in the world, i.e., to infer the cause of a sensory input. First, there is likelihood, “which is the probability of the effect you observe […] given the particular hypothesis you are considering right now” and, second, the “prior probability of the hypothesis (or just the “prior”), which is your subjective estimate of how probable that hypothesis is independently of the effects you are currently observing.” (Hohwy, 2013, p. 17). Taking priors into account also shows that models are highly context-dependent and heavily draw on what the system already knows, meaning that some models may be initially chosen over others because of a system’s current or previous setting. If you have been interacting with your cat more often than your partner in the past hours, your brain will favor ‘cat-associated models’ over ‘partner-models’. In this process, the posterior probability of the hypothesis is generated. The next step now is to compare these posterior probabilities of different hypotheses and find the one with the highest probability. This final step then steers behavior and beliefs.

6.2. Predictive Processing and the Free-Energy Principle

So far, my descriptions of PP have indeed been rather brain-bound. However, when we broaden our scope, we see that PP can be framed as an instantiation of the so-called free-energy principle (FEP). Viewed like that, it become more obvious that there is a deep relation to the very biology of organisms. In this chapter, I will present PP as an instance of FEP and show how such a broader view enables to take seriously the biological reality of agents in a world. This is, importantly, one demand from the phenactive camp to the cognitive sciences, namely to pay more attention to real-life challenges and the biology of cognitive agents.

Karl Friston is the pioneer of describing PP in relation to FEP. As he himself (e.g., 2010, p. 1, 2012, p. 89) contends continuously, the basic idea behind FEP is fairly simple, however, it has manifold and complicated implications. In various publications, he not only shows the implications for perception, action and cognition (e.g., Friston, 2012), but also ambitiously claims that FEP can be seen as a unified brain theory (e.g., Friston, 2010) and can explain why biological systems work the way they do (e.g., Friston & Stephan, 2007). The first connection between PP and FEP is that minimizing prediction error can be interpreted as minimizing free-energy. Also, Friston and Stephan (ibid.) show how FEP, which is an idea stemming from theoretical physics, relates to the way biological organisms and brains work. It is precisely this relation that proves that cognition is profoundly constrained by an organism's physiology. In the following, I will elaborate on these points in more detail.

(1) In chapter 6.2.1., I describe the tasks a biological organism needs to solve in order to sustain its existence. This will further be elaborated on with respect to FEP. The notion of 'embodied inference' is introduced and explained.
(2) The next subchapter 6.2.2. deals with the relation of FEP and PP. In short, minimizing prediction error is a means to reduce free energy and thus to maintain homeostasis, keeping the organism alive.

6.2.1. Keeping Oneself Alive and Well – The Free-Energy Principle

Biological organisms face an ever changing environment and need to maintain their homeostasis. The specific challenges of doing so are different for each phenotype – while cats, to give a simple example, should try and stay on dry land where they find food and can breathe properly, fish would die if they lived under the same conditions as cats. The phenotype of an organism can thusly be defined as a limited set of physiological and sensory states. This state space can be expressed in a probabilistic way; being on land is a more probable state cats would find themselves in, while it would be highly unlikely for a cat to be in water for a long time. This implies that if a system moves out of its expected set of states (e.g., a fish strands at a beach), the state it will be in is then – expressed in information theoretical terms – surprising. It is surprising because the probability that the system will be in that state is very low. Surprise is thus dependent on phenotypes; what is surprising for cats can be a very probable state to be in for fish. In more formal terms, the probability of sensory states must have low entropy, where entropy is a measure of disorder or uncertainty and expresses the average of surprise (cf. Friston, 2010, p. 1).
According to the second law of thermodynamics, the entropy of a closed system increases with time. Biological systems, however, are considered open systems in the sense that they are able to exchange energy and matter with their environment. They are thus able to resist the second law of thermodynamics and sustain their order. How is that achieved? Friston and Stephan (2007, p. 422) suggest that the "premise here is that the environment unfolds in a thermodynamically structured and lawful way and biological systems embed these laws into their anatomy." In this sense, we can talk about embodied systems as *being* models of

the environment they live in (cf. Friston, 2012, pp. 89–90), instead of merely talking about systems that *have* or *build* models of the world. This is what Friston (2012) calls "embodied inference". More specifically, this term expresses that the physiology of a system already presupposes the circumstances it lives in – an organisms's phenotype determines its possible state space.[3]

6.2.2. Predictive Processing as an Instance of the Free-Energy Principle

However, there is a crucial problem for biological systems. They cannot directly minimize surprise, since this would mean to evaluate and consider every possible state they could ever be in at any given time. What is possible, though, is to reduce free-energy and in this way indirectly suppress surprise. Free-energy is considered an upper bound on surprise, which means that it is always greater than surprise. When free energy is minimized, surprise is reduced implicitly, too. It is this aspect that relates FEP to PP.

To see this, keep in mind that a direct evaluation of surprise is not possible. Instead, agents reduce the discrepancy between predicted and actual sensory signals and thus put a lower limit on surprise. This discrepancy is free-energy, which then can be described as the long-term average of prediction error (cf. Seth, 2015b, p. 6).[4] As always, Andy Clark (2013, p. 6) puts this complicated idea into more understandable words:

> Thermodynamic free energy is a measure of the energy available to do useful work. Transposed to the cognitive/informational domain, it emerges as the difference between the way the world is represented as being, and the way it actually is. The better the fit, the lower the information-theoretic free energy (this is intuitive, since more of the system's resources are being put to "effective work" in representing the world). Prediction error reports this information-theoretic free energy, which is mathematically constructed so as always to be greater than "surprisal" (where this names the sub-personally computed implausibility of some sensory state given a model of the world […]). Entropy, in this information-theoretic rendition, is the long-term average of surprisal, and reducing information-theoretic free energy amounts to improving the world model so as to reduce prediction errors, hence reducing surprisal (since better models make better predictions).

[3] A similar thought can be found in the tradition of cybernetics, as Anil Seth (2015b) compellingly shows. He aims to formulate PP from a perspective that combines FEP and cybernetics, showing that Conant and Ashby's (1970) paper pursues a view that is on par with FEP. The title of their work sums up their main claim, viz., "Every good regulator of a system must be a model of that system". The rationale is that in order to handle disturbances coming from external systems, the controller of a system must instantiate a model of this external force (cf. Seth, 2015b, p. 8). Thus, in order to maintain homeostasis, a system embodies and generates a model of its environment.

[4] This statement is a fairly gross generalization and summarizes the core of a mathematically complex matter. For a detailed treatment, see Friston, 2009; Friston & Stephan, 2007.

The ideas I have presented here show a deep connection of biology and cognition. Not only does the kind of mind a system has depend on the kind of its physiological features. Also, biological systems have to mirror (or, 'embody') the causal structure of the world in order to sustain their existence. There are two ways, it seems, this is instantiated. First, recall that generative models are modulated on the basis of error signals, which contain information about the causal structure of the world. This process was described as 'folding' knowledge about causal regularities into the brain's predictions. Secondly, a system's very bodily morphology is structured so to model the system-external world.

These considerations show that although PP puts forth a quite central role of the brain, it still integrates a deep sense of embodiment and relation with the environment in virtue of being an instance of FEP. As such, this theory displays a fundamental continuity of mind and life. Again, there lies a great opportunity to satisfy demands from phenactivism in taking this continuity seriously and exploring its consequences for a theory of the mind.

6.3. Active Inference

What has been described so far is called perceptual inference – the ongoing process of forming models and testing them against sensory input, generating prediction errors and reducing them by correcting the model. The model which best explains sensory data and thus most efficiently reduces prediction error is the model that populates perception (cf. Hohwy, 2012, p. 6). In other words, "[t]he task of perception, given such a multilayered prediction machine, is to match the multimodal sensory signal with apt, consistent, top-down predictions at every level." (Clark, 2014, p. 235)

There is, importantly, another strategy to minimize prediction error, a strategy that reveals yet another deep connection of PP and embodied cognition. Instead of changing models, biological agents can also act so to change sensory input to make it accommodate predictions better. This is called *active inference* and is as essential to minimizing prediction error as perceptual inference.

(1) In chapter 6.3.1., I describe the notion of active inference in more detail and show that it is a crucial part of the overall process of minimizing prediction error.

(2) The role of precision estimations is important when it comes to determining whether perceptual or active inference will serve better to solve a task at hand. In chapter 6.3.2., this role is scrutinized.

(3) The last subchapter 6.3.3. shows that accepting PP eventuates in a revision of well-established models of motor control. The novel PP take on motor control will be important for later chapters which deal with the issue of self-other distinction.

6.3.1. Minimizing Prediction Error by Acting

Instead of merely being the output of some cognitive process, action is an indispensable part of the core mechanism of the brain (i.e., prediction error minimization):

> Action is not so much a response to an input as a neat and efficient way of selecting the next "input", and thereby driving a rolling cycle. These hyperactive systems are constantly prediction their own upcoming states, and actively moving so as to bring some them into being. We thus act so as to bring forth the evolving streams of sensory information that keep us viable (keeping us fed, warm, and watered) and that serve our increasingly recondite ends. (Clark, 2015a, p. 2)

In making this statement, Clark suggests a twofold role for action. First, it fundamentally contributes to the successful prediction of sensory input and thus to the successful navigation of a system's environment. This fits the FEP-formulation quite well, since FEP posits that our actions should minimize surprise, too. Biological systems, as detailed above, are able to interact with their environment so to change its states (and not merely letting the environment change *their* states; cf. Friston, 2009, p. 295). Further, Clark (at least implicitly) corroborates the view I have presented above, namely that PP closely relates to a broader view of a system's cognition in terms of its biology. By arranging sensory input in a way so that it better fits its hypotheses, action profoundly contributes to sustaining a system's order and existence. This line of thinking challenges what Hurley called the 'sandwich' view of the mind, which fosters a clear-cut distinction and division of labor of action, perception and cognition. Much rather, PP seems to imply that the three work in tandem and the lines become more and more blurred.

While PP leaves perception and action computationally equal, it has been claimed that there is an important difference in the so-called direction of fit (Anscombe, 1975; Clark, 2015b, p. 7). Perception changes internal models so they fit the external world (mind-to-world direction of fit), active inference changes the environment so they fit the internal models (world-to-mind direction of fit). However, Madary (2015) convincingly argues that the mind/world distinction needs revision in order to fit PP. Considering that an organism's phenotype is specified by the possible states it can be in, it is proposed that instead of thinking in terms of mind-to-world or world-to-mind directed mental states, it would be more plausible to speak of "world-relative-to-organism-directed" (Madary, 2015, p. 4) mental states. This view takes seriously the connection of PP, action and embodiment, since "[o]rganisms are interested in sustaining their integrity and physical existence; they are interested in what the world is like *relative to their own particular sensorimotor trajectory through the world.*" (ibid.) In this sense, the causal structure that is mirrored by the organism contains the causal structure "relative to the embodiment [...] – and perceptual history – of the perceiver." (ibid.)

6.3.2. Estimating Uncertainty – The Role of Precision Weighting

We have seen that both perception and action are in the business of minimizing prediction error. One question that arises, though, is the following: How does the brain 'decide' which of these strategies to choose? This is where precision expectations or second order statistics enter the picture. The 'job' of prediction errors, as we have seen, is to improve the generative model. However, it is important that errors are not simply taken to be 'trustworthy' (i.e., reliably informing the systems about external states) and change the model or motivate to act. They are taken into account depending on an estimation of how reliable a prediction error will be. If error signals are, for example, estimated as highly precise, they are likely to change the hypothesis. That is, the impact of an error signal depends on how precise it is estimated to be. At the same time, the reliability of predictive models needs to factored in, too. Thus, "we are, in effect, estimating the uncertainty of our own representations of the world." (Clark, 2015a, p. 12)

Precision weighting increases or decreases the gain of error units and thus is also able to change the influence of either top-down prediction or bottom-up sensory information. This also gives us an idea of how the brain 'chooses' which option (perceptual or active inference) will minimize prediction error most effectively. If there is a high-precision prediction error, the predictive model is altered so to fit the sensory income better (perception). If, however, sensory signals are expected to be noisy and unreliable, this will cause the system to actively explore the environment to gather more information (action). It would be unwise to let prediction errors that are deemed inaccurate change the internal model and the safer way here is to amplify the hypothesis.

How the application of precision expectations enables a system to use both fast, simple solutions and more sophisticated strategies is summarized by Clark (ibid., p.12):

> Within that web [of processing], changing ensembles of inner resources are repeatedly recruited, forming and dissolving in ways determined by external context, current needs, and (importantly) by flexible precision-weighting reflecting ongoing estimation of our own uncertainty.

As this quote already gives away, prior knowledge and context play a crucial role in whether or not the sensory signal is deemed reliable. For it may be, in a thick fog, be fatal to rely on sensory sources and more efficient to engage well-probed predictions. In this example, both context (fog) and priors (vision is not a reliable source in a fog) play a crucial role in selecting whether the model or the input is given the higher precision.

It is exactly this selection process that is described as attention. PP systems, as detailed above, take into account their own uncertainty, as well as the noise of external stimuli. We saw that this estimation amounts to the modification of how large the impact of prediction error will be. Thus, this process can be spelled out as a selection process. Former theories of attention have described the phenomenon as specifically that, a process of selection.

To understand the connection, two central notions must be clarified: precision and accuracy. The letter refers to the fact that the more and better prediction error is minimized, the better and more accurate the causal structure of the world is represented. Accuracy thus is a property of so-called “first order statistics” (Hohwy, 2012, p. 4), that is, model states and parameters. Precision, in turn, shows how reliable or precise prediction errors are. Precision expectations thus enable the system to weigh prediction error accordingly (see also above). These are so-called “second order statistics” and show the model certainty.

Precision expectations are highly context- and modality-dependent. While vision is very reliable in clear weather, it may be less so in thick fog. Second order statistics are thus also influenced by prior experience of the system in order to accurately represent what should and should not change predictive models. Both kinds of statistics then together minimize prediction error, where “[…] precision refers to the inverse amplitude of random fluctuations around, or uncertainty about, predictions; while accuracy (with a slight abuse of terminology) will refer to the inverse amplitude of prediction error per se.” (ibid.).

Hohwy goes on to argue that the optimization of precision expectation maps onto attention, since they largely overlap in their functionality. To see this, consider that precision expectations determine whether or not a model should be changed. They thus select the content of perceptual states, in virtue of defining which hypothesis has the highest probability. Furthermore, as Clark (cf. 2016, p. 57) emphasizes, attention selects whether top-down or bottom-up processing should be engaged:

Attention, thus construed, is a means of variably balancing potent interactions between top-down and bottom-up influences by factoring in their so-called 'precision', where this is a measure of their estimated certainty or reliability (inverse variance, for the statistically savvy). Action hence plays a central role when it comes to estimating reliability. Hierarchical generative models encode in which way information must be sampled in order to increase precision (e.g., by encoding saccadic eye movements to scan the scene).

Furthermore, it has been suggested that impairments in attentional processing result in psychiatric or neurological conditions, such as schizophrenia (Fletcher & Frith, 2009), and autism (Hohwy & Palmer, 2014; for an excellent review on this, see Clark, 2016, ch. 7). The basic idea is that the balance of top-down and bottom-up influences plays a major role in bringing forth neurotypical processing and thus 'normal', undisturbed perception. This balance usually helps the system to deal with the vastness of possibilities, potential predictions and signal sources. By weighting competing hypotheses or inputs, those deemed most reliable are privileged and chosen to influence processing. However, once this balance goes astray, in one direction or the other (i.e., by either weighting precision of predictions too high or too low, or weighting precision of prediction errors falsely), pathological symptoms of a great variety occur.

Clark (cf. 2016, pp. 78–79) further contributes to this picture by highlighting that precision expectations may be especially useful to unravel the relation between the sub-personal and personal level. There is an important difference between 'suprisal' and 'surprise', where the former refers to unexpected sensory signals at the sub-personal level and the latter expresses the personal level experience of surprise. The two can come apart when the brain's best way to deal with surprisal still leads to a percept that is highly surprising at the agent-level. Clark exemplifies this by asking the reader to imagine his surprise when a magician makes an elephant appear out of nowhere. For the agent, this is a very surprising experience. However, the winning hypothesis still is that of an elephant in the room, since it best explains away prediction error: "The elephant-on-stage percept is thus the winning hypothesis *given* the current combination of driving inputs, precision expectations, and assigned precision (reflecting, as we saw, the brain's degree of confidence in the sensory signal)." (ibid., p. 78) Thus, although the elephant-hypothesis was initially deemed very unlikely at high-, agent-levels of the hierarchy, it is still precise and powerful enough to overrule the systemic priors that estimated an elephant on the stage improbable at first. According to Clark, this initial un-likelihood is what leads to the agent-level phenomenology of surprise:

> The feeling of surprise, that is to say, might be a way of preserving useful information that would otherwise be thrown away – the information that, prior to the present evidence-led bout of inference, the perceived state of affairs was estimated as highly improbable. (Clark, 2016, p. 79)

What we get here is a neat distinction between the sub-personal and personal level of description, which should not be confused. This relates to my point that inference in the PP scheme is mostly meant as a mechanism that operates at the sub-personal level and has little to do with the conscious elaboration on a specific situation.

6.3.3. A New View on Motor Control

Precision also plays a major role in what we perceive and whether or not we act. To see this, we first have to understand how motor control is interpreted within the PP framework.

There is one rather radical consequence of spelling out action in terms of active inference, a consequence that complements, but also challenges the well-established theory of motor commands and efference copies for motor control (Clark, 2016; Pickering & Clark, 2014).

A famous example of such a theory is the HMOSAIC (Hierarchical Modular Selection and Identification for Control) account as formulated in Wolpert et al. (2003). Motor commands are here used to execute desired movement trajectories (intentions). In more detail, an intention serves as input to an inverse model which then outputs the motor command. This commands is sent to the motor plant – the body – which implements the movement. The sensory feedback of the motion is fed back and used to update the desired movement trajectory. In this way, the motor command is updated and optimized, since the feedback provides information about discrepancies between motor command and actual trajectory. In order to make this process faster and to avoid errors, a so-called efference copy is sent to the forward model, which estimates sensory feedback without actual execution. This non-actual sensory feedback is then used to update the trajectory and is also compared with the actual feedback, so to improve the forward model.

To see how this model is different from the PP solution, consider that active inference is yet another way to minimize prediction error. Motor control here is nothing else but top-down prediction of sensory consequences, i.e., which sensory input would occur if my body moved in a certain way. If no movement occured although it was predicted that it would, a large prediction error occurs. This proprioceptive prediction error may count as a motor command, since it is used to change the state of muscles – the body moves according to the prediction.

In this sense, action minimizes proprioceptive prediction error, because it helps fulfill proprioceptive predictions. There are two important implications here.

First, precision estimations once more become central. If, for example, upwards flowing prediction error was deemed highly precise, it would be powerful enough to actually change the prediction from 'moving' to 'not moving' and no movement would occur. Precision thus needs to be attenuated for self-generated movement (cf. Pickering & Clark, 2014, p. 454). Secondly, and this is more closely related to the current topic, the PP account renders inverse models redundant, since the forward model is already doing the job of it, viz. predicting sensory consequences. This complex forward model thus not only predicts, but also brings forth movement trajectories.[5]

We now see how precision determines action; only when a balance between top-down and bottom-up influences is achieved can action ensue. When sensory prediction errors are not attenuated, no action occurs. In turn, when predictions are not estimated reliable enough, prediction errors (i.e., sensory signals that no movement is happening) can change them and again inhibit movement (Fig.4).

There are consequences for the sense of agency of the account just depicted. It has been hypothesized that the discrimination between internally and externally generated movements draws on the attenuation of sensory prediction error. Self-generated movements need attenuation of prediction error, as we have seen, to make movement ensue. Thus, the sensory consequences of internally generated movements are suppressed. When movement is induced externally, however, predictions about sensory consequences are much less accurate, thus not able to suppress prediction error. Externally generated movements elicit stronger proprioceptive prediction errors. This will be crucial when considering a functional self-other distinction in a later stage of this thesis, for it is claimed that

> [f]orward models of action and the corollary discharge are thought to be crucial in determining ownership of action, or sense of agency, and being able to distinguish between self and other by distinguishing between self-generated actions and movements generated by external forces (Brown & Brüne, 2012, p. 5)

[5] Pickering and Clark (2014, pp. 455–456) contrast these two account and propose different ways to empirically test which of them is more promising.

What this quotation shows is that a PP take on motor control in general also provides insights on a fundamental component of social cognition (i.e., self-other distinction). I will elaborate on this idea in chapter 7.3.

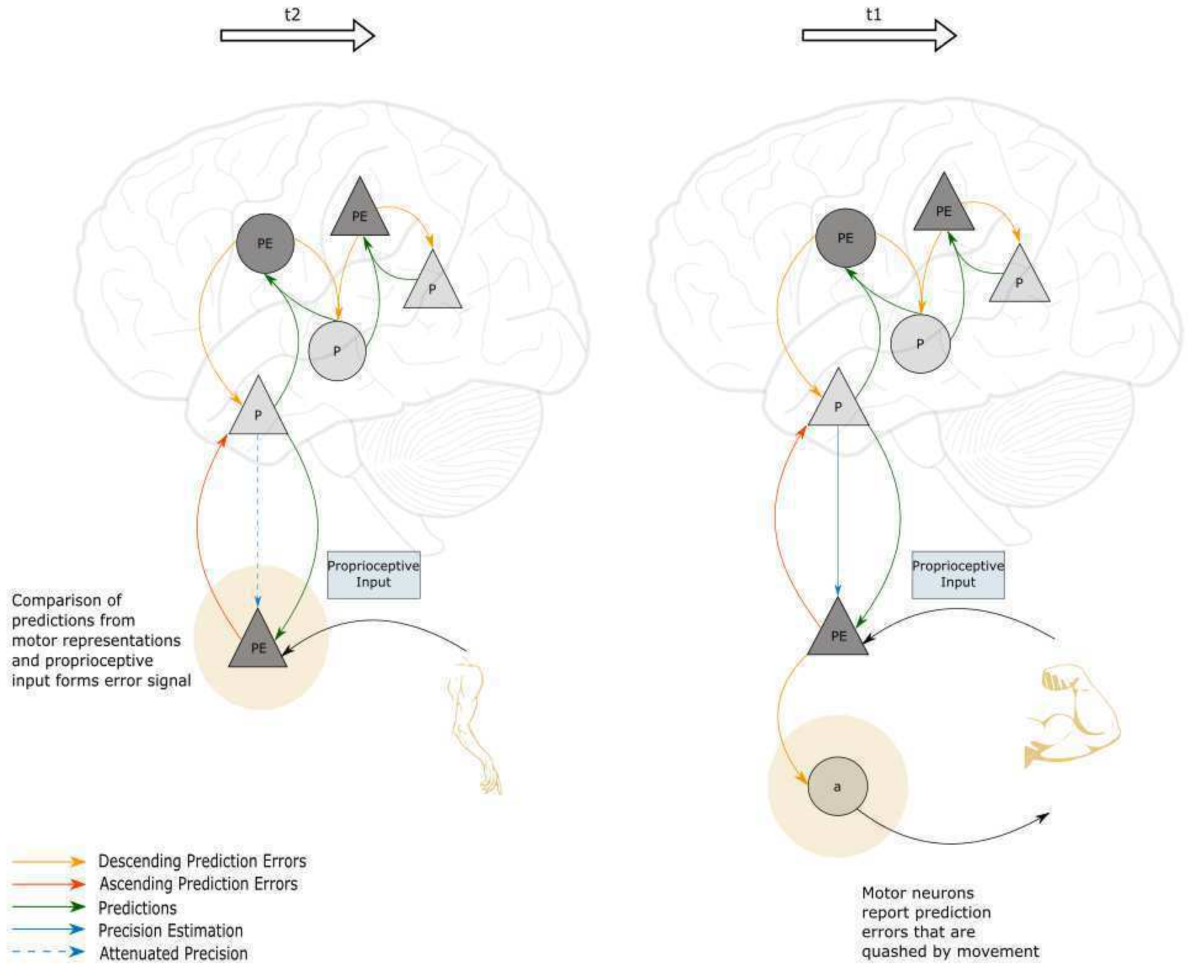

t2
t1
PE
P
PE
P
P
PE
Proprioceptive Input
Comparison of predictions from motor representations and proprioceptive input forms error signal
PE
P
PE
P
P
Proprioceptive Input
PE
a
Motor neurons report prediction errors that are quashed by movement
Descending Prediction Errors
Ascending Prediction Errors
Predictions
Precision Estimation
Attenuated Precision

Fig. 4 Active inference

This figure depicts a rough sketch of the dynamics of active inference. At time t1, predictions (light grey units 'P') about the sensory consequences of movement are passed down via backward connections (green arrows) and compared against actual incoming proprioceptive input. Since predictions and signal conflict in that movement is predicted, but has not yet occurred, a prediction error (grey units 'PE', light orange circle) is generated. Note that precision of this ascending error signal (red arrow) must be attenuated (dashed blue line), so not to change predictions and thus to inhibit movement. At time t2, motor neurons in the spinal cord (a) report prediction errors that are quashed by movement. In this way, predictions about movements are fulfilled by eliminating proprioceptive prediction error.

6.4. PP and Interoception

PP has mainly been described as a principle for exteroception, perhaps because of its historical roots in a Helmholtzian framework and its early adaptations to (mainly) the visual system (Tsakiris, 2008). However, as has been remarked by Seth (2015b), in order to keep itself alive, a system not only has to infer exteroceptive and proprioceptive signals. Knowing and steering its interoceptive states is just as important and fundamentally contributes to an organism's well-being.

In order to understand why an extension of PP to interoception is needed, consider that from the perspective of the brain, the body is just as 'hidden' as the external world. This means that in the same way as the brain cannot directly access environmental states, it cannot directly access the states of the body it is embedded in. It needs to infer those just as it needs to infer hidden causes in the outside world. A dangerously low blood sugar level or a too fast heart rate could be potential dangers that an organism needs to be able to regulate. According to Seth (2013, p. 566), maintaining homeostasis is achieved by so-called interoceptive inference (or interoceptive predictive coding): "From a PC [predictive coding] perspective, this implies that an organism should maintain well-adapted predictive models of its own physical body (its position, morphology, etc.) and of its internal physiological condition." In this chapter, I introduce the broader topic that Seth embeds this claim in and then elaborate on the notion of interoceptive inference and its consequences for the scope of PP:

(1) Chapter 6.4.1. couches the idea of PP in a cybernetic perspective, as proposed by Seth. This results in a fruitful application of PP to interoception, hence interoceptive predictive processing (IPP).
(2) In the following chapter, it is described in more detail how IPP is proposed to work. It will become clear that subjective emotional states are thought to arise throughout the process of active inference, that is, the update of predictions of a specific physiological state.

6.4.1. The Cybernetic Brain

In a recent paper, Seth argues that instead of seeing the roots of PP in a Helmholtzian framework of perception (e.g., Friston, 2009), it is more useful to relate PP to cybernetics.

In doing so, it becomes obvious that PP not only applies to exteroceptive and proprioceptive inference, but also the inference of interoceptive signals. Without going into much detail, let me briefly summarize this claim.

The basic assumption that relates cybernetics to PP is that both assert that a system strives for stability and in doing so models its environment. It does so, according to cybernetics, by implementing two levels of feedback: “a first-order feedback that homeostatically regulates essential variables (like a thermostat) and a second-order feedback that allostatically re-organises a system’s input-output relations when first-order feedback fails, until a new homeostatic regime is attained.” (Seth, 2015b, p. 7) This directly relates to FEP, since FEP states that in order to keep itself alive and in order, the system instantiates or embodies a model of the environment. The quotation also shows that a system needs an additional circuit of feedback in order to regulate its homeostasis.

As is also mentioned by Seth, connecting PP and cybernetics brings up a key challenge. While the former is a deeply representational theory, proponents of the latter emphasize that no such things as representations in a strong, explicit sense are needed. In a cybernetic picture, homeostasis can be maintained by an implicit, embodied modelling of the world. On the one hand, Seth (ibid., p. 9) allows such a view, but also argues that representations and inner models are still the best way to refer to what a system does:

> However, where there exist many-to-many mappings between sensory states and their probable causes, as may be the case more often than not, it will pay to engage explicit inferential processes in order to extract the most probable causes of sensory states, insofar as these causes threaten the homeostasis of essential variables.

To describe a possible connection between PP and cybernetics serves two purposes here; first, it displays an even more intimate relation of mind and biological life and makes sense of the claim that agents embody the world. This has been described in the last paragraph. The second purpose is to emphasize that for biological agents, it is not only important to map and infer exteroceptive signals, but that it is just as crucial for sustaining their existence to monitor their inner, that is, *interoceptive* states. What emerges here is a more comprehensive picture of PP, since it now also implies the inner environment, that is, the body of agents.

6.4.2. Interoceptive Predictive Processing

Seth goes on to draw a picture of interoceptive inference, that is, the suppression of interoceptive prediction errors in comparison to top-down predictions of internal states. In general, interoceptive predictive processing (IPP) follows the same principle as inference of exteroceptive or proprioceptive states as described above.[6] Predictive models of the next most probable interoceptive signals are generated and compared to actual input. According to the size and estimated reliability of the prediction error, there are (at least) three different ways to minimize prediction error; namely

> by (i) updating predictive models (perception, corresponding to new emotional contents); (ii) changing interoceptive signals through engaging autonomic reflexes (autonomic control or active inference); or (iii) performing behavior so to alter external conditions that impact on internal homeostasis. (Seth 2015b, p. 10)

There are several important implications in this quotation. First, the engagement of autonomic reflexes serves to retain or maintain homeostasis. This directly relates to the cybernetic perspective described previously. Throughout the process of monitoring its own states, the system keeps updating and changing its states, so to keep itself in equilibrium. The performance of behavior has the same goal, as is obvious in the quotation. The second important point made here is that the process of updating interoceptive predictive models brings forth emotional contents: "emotional content is generated by active 'top-down' inference of the causes of interoceptive signals in a predictive coding context." (Seth, 2013, p. 565)

This model of emotions deserves some attention. For it draws on the famous James-Lange theory, which assumes that emotional experiences are first and foremost constituted by the monitoring of interoceptive states (James, 1950). However, one shortcoming of this theory, which has been adopted and refined widely (e.g., the somatic marker hypothesis, Damasio, 1996), is that it (at least implicitly) asserts a one-to-one mapping of emotional effect and interoceptive cause (cf. Clark, 2016, p. 233). This means that it is assumed that there is one change in interoception (that can probably be marked, hence 'somatic *marker* hypothesis') that leads to the perceived emotion. However, it is more likely to assume that there are several possible causes that bring forth an emotion.

Viewing the basic idea of the theory within an IPP perspective, though, promises to alleviate this problem of former theories in the sense that the most probable of a variety of likely reasons is actively inferred. IPP furthermore assumes that hierarchically higher levels

[6] For a proposal on how IPP might be implemented in the brain, see Suzuki et al., 2013.

integrate exteroceptive, proprioceptive, and interoceptive information and generate predictions of probable interoceptive states. In this sense, Clark (2016, p. 234) explains:

> A single inferential process here integrates all these sources of information, generating a context-reflecting amalgam that is experienced as emotion. Felt emotions thus integrate basic information (e.g., about bodily arousal) with higher-level predictions of probable causes and preparations for possible actions.[7]

In other words, emotions occur in the process of integrating information from several sources in order to form interoceptive predictions. These predictions, which then contain contextual information from exteroception, interoception and proprioception are compared to actual incoming signals. Throughout this process, emotional states arise. Note that this implies an intimate, probably interacting relation between interoceptive and exteroceptive processing. For one, both external and internal events shape emotional response. On the other hand, when external events trigger a cascade of interoceptive processing, these events will be "affectively coloured" (ibid.).

The role of the insular cortex, especially the anterior insula (AI) has been proposed to play a key role in this process. Seth (cf. 2013, p. 568) claims that emotional responses crucially rely on predictions of causes of interoceptive input that are continually updated. These predictions are generated, compared to actual input and then updated within a salience network that consists of AI, ACC and several other functional connections to the brainstem (Fig. 5)

This IPP view perspective further supports a prominent view on how a sense of self and body ownership is brought forth (e.g., Blanke & Metzinger, 2009; Tsakiris, 2008). According to this perspective, representations of the self are basically body representations, integrating multi-modal information and exteroceptive stimuli. The relation between IPP and this idea becomes obvious in Apps and Tsakiris' (2014, p.88) statement that a sense of self arises when the causes of signals that are 'most likely to be me' are inferred:

> That is, one's own body is the one which has the highest probability of being "me" as other objects are probabilistically less likely to evoke the same sensory inputs. This information can be considered as highly abstract with respect to the low-level properties of the stimuli

[7] One important consequence of adopting such a view is that it makes 'two-factor' theories of emotional experience obsolete. These theories suggest that emotions have two components: a bodily and a cognitive one. The latter is needed to appraise the former and thus leaves the emotion subjectively colored. However, IPP now claims that subjective emotional states simply occur in virtue of the single mechanism just described. For a more detailed discussion, see Seth, 2011.

> and can only be represented as “self” when different streams of multisensory information are integrated.

What the authors suggest here is that only the integration of a multitude of information from different sources brings forth a sense of self and body ownership. This idea is able to explain phenomena that hint to a plasticity of phenomenal selfhood, such as the rubber hand illusion (RHI) or the enfacement illusion (Tsakiris, 2008).

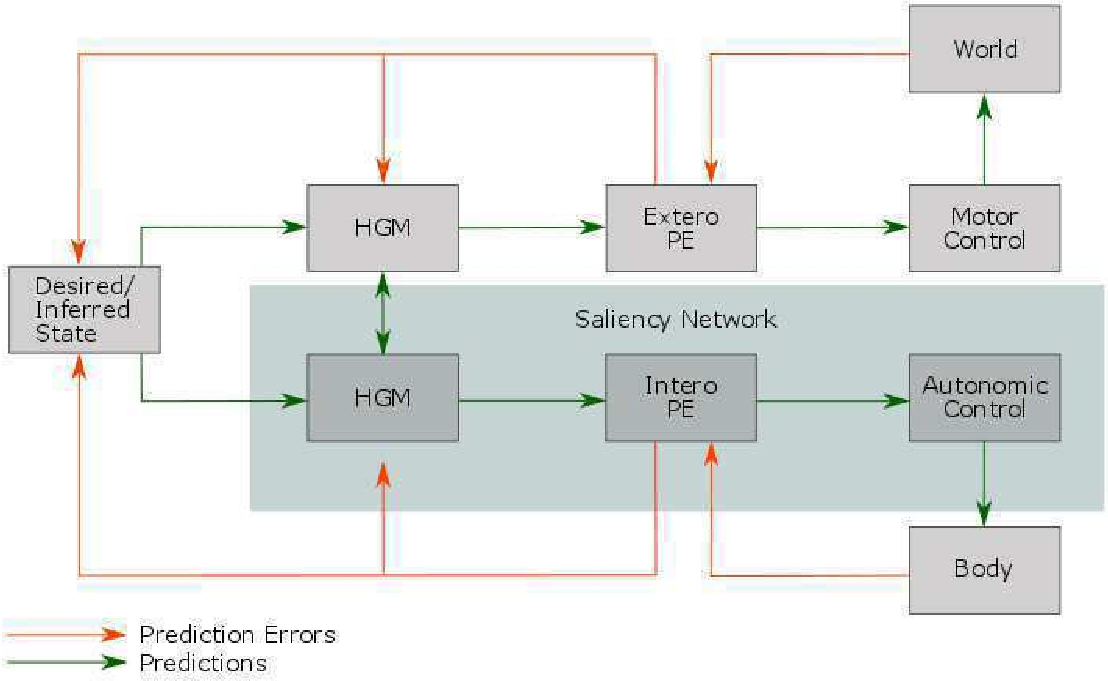


Fig. 5 Interoceptive predictive processing

Emotional states are proposed to depend on updated hierarchical generative models (HGM). On the basis of a desired or inferred state of an organism, predictive models of interoceptive and exteroceptive information are generated. Prediction errors are used to engage either classical reflex arcs (i.e., motor control) or autonomic reflexes (i.e., autonomic control). This process is called active inference and serves to change states of the world or the body. The resulting error signal is used to update not only the generative model, but also the desired or inferred physiological state. Interoceptive predictive processing is realized by the salience network. Figure adapted from Seth, 2011.

In these illusions, external objects (e.g., a rubber hand) or the face of another person are manipulated in such a way that a good amount of participants experience them as ‘being my own’:

> In the RHI, subjects watch a rubber hand being stroked, while their real hand is stroked synchronously. A good amount of subjects experience the increasing sense that the rubber hand becomes their own hand. The reason for that, or so it has been claimed, is that visual

> and tactile inputs overrule conflicting proprioceptive information. This means that – from a PP perspective – the experience of the rubber hand belonging to my body is brought forth by an update of empirical priors about which hand is mine, due to multisensory prediction errors. The likelihood that my real hand is my own hand is diminished, while the probability that the rubber hand is my own increases (cf. Apps & Tsakiris, 2014, p. 90).

What speaks for a contribution of IPP to these processes is that in the RHI, the temperature of the actual hand decreases, potentially showing interoceptive active inference (Seth, Suzuki & Critchley, 2011, p. 4). There is further evidence that indirectly speaks for the involvement of IPP in body ownership illusions. Tsakiris and colleagues (2011), for example, found a correlation between susceptibility to RHI and interoceptive sensitivity, which was measured using the heartbeat detection task.

The idea of IPP and its role for phenomenal selfhood will proof important for explaining aspects of social cognition, too. In chapter 7.3., I will scrutinize how the process of multisensory integration – which also implies integration interoceptive and exteroceptive information – may provide a very basic means to distinguish self from other.

6.5. A New View of the Mind

So far, I have drawn together interpretations of PP that depict action, cognition and perception as both representational while at the same time deeply connected to brain-external structures. Both Seth (2015a, 2015b) and Clark (2013, 2015a, 2015b) emphasize that PP fosters a renewed picture of the mind, and more specifically of representations. There are several ways in which this new notion departs from a more classical picture, which I will present in this section:

(1) First, I present Seth’s notion of counterfactually-equipped representations in section 6.5.1. This view not only offers to integrate a compelling idea from the phenactive camp, but also lays the groundwork for discussing the phenomenological level of description in later chapters.

(2) In chapter 6.5.2., I discuss how PP changes the concept of representations in general. Following Clark, I claim that it yields a way to get rid of a radical cognitivist view on representations and instead puts forth an interpretation of representations as abstractions. Further, it is possible to describe them in the PP framework as profoundly ‘action-oriented’.

(3) I then turn towards PP's unifying power and show that it indeed integrates important insights from both phenactive and cognitivist camps.

(4) Finally, in chapter 6.5.4., I claim that PP and 1-3E are a perfect match and show how they can be applied to each other. I submit that 1-3E can be seen as yielding conceptual heuristics for PP and thereby establishing a basis for the application of both theoretical building blocks to social cognition.

6.5.1. Counterfactually-Equipped Representations

In Seth's (2014) theory of PPSMC (predictive processing of sensorimotor contingencies), he shows how so-called counterfactually-equipped hierarchical models can be related to a phenactive view of perception; the theory of sensorimotor contingency. Although the latter is an anti-representationalist theory, Seth successfully distills its most important claims and connects them elegantly with PP, embedding the ideas of sensorimotor contingencies with a representationlist view. In the sensorimotor theory of perception (SMT; Noë, 2004) it is claimed that we do not perceive the world by *representing* it, but that perception arises in virtue of learning how to *act* on the world. SMT asserts that we have implicit knowledge about how our perspective on an object would change if we were to move, thus gathering additional sensorimotor knowledge about this object. Behind this approach stands the question of how we perceive objects as wholes if we only see one side of them. If, for example, you see your cat lying on her stomach, you cannot *see* her stomach (nor touch it, God forbid!), but you still *know* it is there. This is so, according to SMT, because you obtain knowledge about *sensorimotor contingencies* (SMC), that is, laws which entail information about how perception changes due to possible motor actions. To know these laws is to master SMCs; in the words of O'Regan and Noë (2001, p. 943):

> The sensory modalities, according to the present proposal, are constituted by distinct patterns of sensorimotor contingency. Visual perception can now be understood as the activity of exploring the environment in ways mediated by knowledge of the relevant sensorimotor contingencies. And to be a visual perceiver is, thus, to be capable of exercising mastery of vision-related rules of sensorimotor contingency.

With a clever twist, Seth (2014) shows us how PP directly relates to SMT and is thus able to pick up its core idea and put it in representational, yet embodied terms. First, remember that active inference can serve different purposes. It can either lead us to act, or help to improve perceptual experience. In the latter case, it may confirm or disconfirm a current hypothesis, or it may help to disambiguate competing hypotheses. This implies that generative models not only encode the most likely cause of a sensory event, but that they contain *counterfactual* information, too. That is, they encode different, hypothetical scenarios, not all of which will actually be executed (cf. Seth, 2015b, p. 19). Such *counterfactually rich* generative models are able to capture the idea of SMCs, that is, how would an object look like *if* a specific action were executed on this object? This is what Seth (2013, p. 19) expresses with his theory of *predictive processing of sensorimotor contingencies*:

> […] a counterfactually-rich hierarchical generative model explicitly encodes probabilistic representations of the external causes and expected values and precisions of fictive sensations conditioned on a repertoire of possible actions, thus capturing the key notion within sensorimotor theory of somehow perceiving parts of an object not directly available within the ongoing sensorimotor flux.

This leads to a renewed concept of representations as counterfactually rich, and in yet another sense involving sensorimotor information, thus not being void of any grounding relation to their embodied roots.

Such a notion of representations will proof to be fruitful for a theory of social cognition. I will detail how the idea of counterfactually-equipped generative models can shed light on social processing in chapter 7.4.

6.5.2. The War on Representations

Clark offers several points in favor of the idea that the kind of representation that PP refers to is in no way related to the stiff, passive-mirror-of-nature representation old-fashioned cognitive science talked about. First, although internal models are a central part of the account, these models are fundamentally grounded in embodiment, in that they “allow a system to combine a real sensorimotor grip on dealing with its world with the emergence of higher-level abstractions that (crucially) develop in tandem with that grip.” (Clark, 2014, p. 242) Representations or internal models are not marooned from brain-external matter, they are *for* engaging the body and world, to elicit action and active navigation of the environment. Again, remember Madary’s claim that an agent’s mental states always depend upon her sensorimotor trajectory and her bodily structures. This brings us even closer to the idea that representations – if PP and FEP are correct – are what Clark (2015a, p.4) has called “action-oriented through and through”.

In that sense, Clark (2015b, p. 3) predicts that PP will bring peace to "the smoking battleground of the Representation wars." However, the concept is not given up. To see this, allow me to cite Clark's (ibid., p.5) idea at length:

> […] each PP level (perhaps these correspond to cortical columns – this is an open question) treats activity at the level below as if it were sensory data, and learns compressed methods to predict those unfolding patterns. This results in a very natural extraction of nested structure in the causes of the input signal, as different levels are progressively exposed to different re-codings, and re-re-codings of the original sensory information. These re-recodings […] enable us, as agents, to lock us onto wordly causes that are ever more recondite, capturing regularities visible only in patterns spread over space and time. Patterns such as weather fronts, persons, elections, marriages, promises, and soccer games. […] What locks the *agent* on to these familiar patterns is, however, the whole mutli-level processing device (sometimes, it is the whole machine in action). That machine works (if PP is correct) because each level is driven to try to find a compressed way to predict activity at the level below, all the way out to the sensory peripheries. These nested compressions, discovered and annealed in the furnace of action, are what I [...] would like to call "internal representations".

As I read Clark, the essence of his claim is that representations are basically abstractions. They are not the sensory data itself, but information that has been compressed and abstracted, a prediction of what the sensory data a level below could be. In this sense it is useful – or so I think – to talk about internal models and representations. Predictions *represent* potential actual sensory input, becoming more and more abstract as one goes up the hierarchy. At this point, it becomes obvious how PP can serve as a background theory for 1-3E and 1-3sE. The notion of representation as laid out above is extremely flexible, thus allowing for the presence of inner models at *every level* of the hierarchy. Even at the very low-levels, one would find representations – namely those units that predict the activity of sensory modalities. Representationalism then arises as a graded phenomenon (cf. Metzinger, 2014a, p. 278) whose degree of abstraction increases. Applied to 1-3(s)E, this means that there would be representational processes at each level of embodiment that are differentiated by their degree of abstraction from the actual sensory input.

These kinds of representations do not merely generate a picture of the world in our heads. If active inference and FEP are taken seriously, they engage the whole agent to extract hidden causes in the world. In this sense, Clark (2016, p. 133) opts for talking about "action-oriented"-predictions: "They will represent how things are in a way that, once suitably modulated by the precision-weighting of prediction error, also prescribes (in virtue of the flows of sensation they predict) how to act and respond."

Considering the role of internal models, namely to prepare systems to act upon their environment and enable them to do so thus helps to tune the notion of representation towards a more embodied, flexible one. This is, in my view, yet another step towards finding the 'golden middle' between cognitivist and phenactive theories.

6.5.3. Phenactive Predictive Processing?

Another crucial point to understand why PP so naturally integrates important claims from the phenactive side of the theoretical spectrum is that the picture of the brain's function and mechanisms is incomplete if the role of the body and action are ignored. For biological agents do have a body and have to navigate their environment with respect to their phenotypical specificities.

Clark (2016) argues in length that the brain's task is to prepare an organism for exactly this bodily exchange with the environment. Importantly, another fundamental part of hierarchical modelling and prediction is to estimate which channels give an agent the best grip on the world. In other words, the brain's function is not only to predict its next states, but also to choose in which way this may be achieved most efficiently. Which action or movements brings forth the manipulation so they best fit the prediction? Is it better to disambiguate incoming stimuli by active or perceptual inference? Clark's interpretation of Friston's take on FEP entails that organisms strive to reduce free energy by opting for the most efficient way. Efficiency, here, means to find the strategy that involves the least complex route, but brings the largest effect. If I was to estimate where a ball lands so I can run and fetch it, it may be most efficient to start running and let my body most of the work. Standing still and computing a rich inner representation of the movement trajectory and then start running would not only be too slow, it would also be too complex for the task at hand.[8]

If such a view of the brain's function is correct, it is obvious that action and embodiment are as fundamental to PEM as is the construction of inner representations. If, or so I shall argue, the goal is to find a theory which helps us to explain and predict real-life phenomena and which enriches empirical research, such a theory is highly attractive. However, it may not be so easy. For it has been argued that extra-neural events may be *enabling*, but not *constitutive* elements. Could it not be that it was enough to stimulate our brains in the right way and bring forth the exact same experience of being what we are, without the need to engage the body or environment?

[8] This has been called the Outfielder Problem (cf. Clark, 2016, p. 190).

This worry has been expressed in the famous ‘brain in a vat’ thought-experiment: What if our bodies, our actions, our environment, and every other person we know are just very convincing phenomenal experiences, brought forth by the stimulation of our brains, which in reality are disembodied meatballs in vats, manipulated by some evil scientist? There always seems to be this last bit of skepticism, which is meant to steer our attention to the possibility that all what we experience may not require a body at all.

This problem has been discussed at length elsewhere (Thompson & Cosmelli, 2013), but I still wish to defend my conviction that a useful theory of (social) cognition should take embodiment seriously. I think it is important to sort out what the brain in a vat scenario really is about. As I see it, this thought experiment targets *conscious experience*, and not *cognition*. While embodiment may not be causally relevant to consciousness (Anderson, 2014), the case may be different for cognition. PP actually supports such a dissociation. To see this, consider that it has been argued that the main task of the cortex is to generate predictions about incoming stimuli. This means, basically, that the brain is able to reconstruct “the sensory signal using knowledge about interacting causes in the world” (Clark, 2016, p. 85). Once learned, the system will be able to process without actual input and thus bring forth imagination, dreams, or mindwandering. In this sense, one could say that conscious experiences have their minimally necessary counterparts within the skull. Cognition, on the other hand, seems to ‘shamelessly’ exploit the body and its environment whenever this is the best way to go.

This leaves us with the following picture of the (conscious) mind. PP accounts for quite a spectrum of phenomena; on the one hand, it is a rather brain-bound view, since the generation of predictions and the precision weighting process is neurally implemented. In that way, perception is brought forth mainly by top-down processing and is determined internally. This side of PP also neatly accommodates ‘representation-hungry’ processes like imagination and dreaming, which seem to occur without much brain-external help. On the other hand, even those more ‘decoupled’ phenomena have been shown to involve the body. Saccadic eye-movements, for example, may be the bodily ‘grounds’ for phenomenal experience in dream states (Metzinger, 2014a).

PP also naturally bonds with claims from embodied cognition and phenactivism. It seems that although a great part of the prediction error minimization machine is located in the brain, body and action play an indispensable role for this mechanism. To see this, remember that while prediction generation clearly is the brain’s job, the minimization of prediction error –

the *core* of PP – heavily engages the body and world in virtue of active inference. It thus seems obvious that embodiment is fundamental to the internal mechanisms in at least two ways. First, the very morphology and its phenotype of a system set the baseline of what are probable states for it to be in. Second, as described above, active inference appears as a part of PP it cannot do without. I thus propose that the *whole* agent should be considered as the prediction error minimization machine.

6.5.4. Merging 1-3E and PP

At first sight, it appears intuitive that 1-3E and PP make a good match. Both assign a fundamental role to representations and models, both argue for a bodily grounding of representations, both present a functional hierarchy.[9] However, for all of these three seemingly common elements, important questions arise that need clarification. First, the notion of representations and models should be scrutinized to ensure a common usage that fits both theories. Further, we need to make sure how and whether PP can be seen as a grounding theory, a property which Metzinger (cf. 2014a, p. 277) explicitly assigns to his theoretical view. Most pressing is then to clear up how levels of embodiment in the 1-3E framework relate to hierarchical levels in the PP architecture.

Although this will only be resolved by the end of this section, let me briefly state one common feature that should suffice until we can tackle the issue in more detail. Very generally speaking, low levels that exhibit processes operating on fast and short spatiotemporal scales relate to the lowest level of embodiment 1E. More abstract and slower processes can then be connected to 2E, where one would also find a certain degree of multisensory integration. Of course, the level of 3E is the trickiest, given that it is still controversial how exactly representational content becomes available for conscious awareness. However, it should be safe to say that 3E entails processes that operate on long and large spatiotemporal scales, thus relating to high levels in the PP architecture. After this very short clarification, I will now proceed to tackle each of the issues named above, starting with discussing the notion of representations in the two theories.

PP is a representational theory, which at the same time assigns a fundamental role to the body and possibly organism-external elements. Similarly, 1-3E puts much weight on

[9] Recall that I refer to those versions of PP which foster embodiment, such as Clark, 2016; Seth, 2015b; Friston & Frith, in press.

representations when it comes to the levels of 2E and 3E. However, the exact properties of the lowest level 1E remain rather obscure, as I have argued in chapter 5.2.3. It appears that 1E does not necessarily entail representations, since it may also describe processes which directly exploit an organism's morphology. For both theories, thus, one needs to answer the following question: What kinds of representations are to be found in an agent's cognitive hierarchical architecture and at which levels are they to be found?

There are two ways to tackle this question. First, regarding Friston's term 'embodied inference', the concept of a 'model' appears in new light. That is, at a very low, biological level, one can claim that the agent is a model of its environment, in virtue of their anatomy which is attuned to the causal structure of the environment. In this sense, the way in which PP enables to talk about 'models' is in a meaningful way stretched and applied to the corporeal area. The notion of embodied inference also smoothly applies to the level of 1E, since it describes how the very morphology of an agent incorporates (and thus models) the laws of nature it needs to follow to ensure survival. Secondly, it is important to recall that the concept of a representation that results from a specific interpretation of PP allows for a quite flexible usage of the term. Representations can – or so I have argued – be seen as abstractions from the actual sensory source. There is no commitment to a specific format of representations as, for example, rule-based, symbolic and explicit computations. To the contrary, representations have been described as 'action-oriented', and also as becoming more and more abstract as one goes up the functional hierarchy. This means, in turn, that at the low end of the hierarchy, we may find very crude and concrete representations that only show little abstraction from the actual sensory signal. Accordingly, Palmer, Seth and Hohwy (2015, p. 378) describe the lowest level of the PP hierarchy as "the sensory input itself – corresponding to retinal or lateral geniculate nucleus activity in the visual system, for example."

Taken together with my first point, we may say that at the lowest level of the hierarchy – the actual sensory modalities, the actual sensory input, and the actual morphology of an organism – the meaning of a 'model' of the environment is fundamentally related to the concrete biology of that organism. At the same time, as one goes up the hierarchy, representations of these elements become more and more abstract. In this way, PP fits the hierarchy depicted in 1-3E. Although in a much more detailed and fine-grained manner, they both describe the same kind of hierarchy.

Now, why do we need 1-3E, if PP is the more fine-grained description? I would argue that the application of 1-3E puts PP in heuristics we are more used to, namely levels of description which enable to talk about the same target phenomenon at different stages of its processing, and thereby makes it an endeavor accessible for interdisciplinary research. The three levels of embodiment can be seen as marking crucial processing stages in the hierarchical architecture whose principles are detailed by putting them into PP terms. Thus, there is no one to one mapping between levels of embodiment and PP levels. Much rather, as the spatiotemporal scale of PP levels changes, levels of embodiment change, too.

We may now start to tackle the remaining issue, viz., the clarification of grounding relations in this newly emerging framework. It is often said that representations are grounded in sensorimotor processes, and the two theories here are no exception. What exactly does this mean, though? To begin with, consider the claim that representationality is a graded phenomenon, meaning that it is probably not an all-or-nothing, but gradually arising process. Now recall that we said that at the lowest level of the hierarchy, there are actual sensory signals, actual exteroceptive, interoceptive and proprioceptive signals. At the next highest level, according to PP, there already exists an abstraction of these signals, namely a predictive, probabilistic model of what this input will be. These abstractions we have called representations and they are grounded in the actual signal in virtue of being corrected by prediction error, carrying information about the external world and thus providing a 'grip' of that world. This is one way in which we may describe representations as grounded and gradual; going up the hierarchy, the degree of abstraction from the actual signal increases, however, even high level predictions are only 'useful' when prediction error is minimized at each level of the hierarchy, thus it appears that each prediction receives a 'real world check' when it finally reaches the lowest level of the hierarchy.

Another meaning of a grounding level emerges when we consider the body as a hyperprior.[10] Hyperpriors basically provide very general and abstract 'priors for priors' (cf. Hohwy, 2013, p. 71). An individual's body can be said to form such a prior because it enables and constrains its sensorimotor trajectory. Whether or not, for example, an individual needs to

[10] Saying that the physiology of an agent is a hyperprior can appear contradictory, since priors and hyperprios are abstract representations operating on large spatiotemporal scales. However, if one considers the notion of empirical priors which I have clarified above, it becomes clear that they are themselves subject to both bottom-up as well as top-down control. As such, even our concrete body can end up as an abstract hyperprior.

be constantly in salty waters to ensure survival necessarily determines its behavior. The types of organisms who do need to be in salty waters have to steer their behavior in a way so to stay in salty waters, or else they cease to exist. This ties back to the notion of embodied inference and describes the fact that an individual's body mirrors the laws of nature it needs to follow to ensure survival and in this sense also determines this individual's action, perception and cognition. Bodily priors can be seen as constrains that dictate further processing and it is quite obvious why these priors – operating on high levels – must be grounded; they contain information about the actual body, gathered through prediction error minimization at *every* level of the hierarchy through movement (active inference). Predictive models of our bodily features must somehow represent our actual morphology in relation to other objects or individuals in order to be useful. They become useful, it appears, by constantly being compared to and corrected by actual sensory signals and in this sense are grounded in the sensorimotor trajectory of an individual.

In sum, it appears that PP and 1-3E indeed are a good match. I showed that there are basically three elements that have been worked out in merging the theory. These were a novel way to view representationality as a gradual phenomenon, the proposal to use levels of embodiment as heuristics to mark stages in the PP hierarchy, and the claim that representations are grounded in sensorimotor processing. In the next chapter, this merger shall be applied to the study of social cognition.

7. Towards a Unifying View on Social Cognition: Levels of Social Embodiment[1]

> *The social domain is both highly complex (frequently involving the appreciation of perspectives upon perspectives, as when we know that John suspects that Mary is not telling the truth). It is, moreover, a domain in which context (as every soap opera fan knows) is everything and in which the meaning of small verbal and non-verbal signs must be interpreted against a rich backdrop of prior knowledge. (Clark, 2016, p. 225)*

In this chapter, I will introduce my own proposal on how to frame social cognition. I will do so, as already indicated, by adopting Metzinger's (2014a) theory of 1-3E with the modifications taken in the previous chapters. As I will show, it will also be very helpful to further embed 1-3sE in the framework of PP. To briefly recapitulate, I have argued so far that a pluralistic perspective is a fruitful way to frame social cognition, but that there are several shortcomings that need to be overcome. The most important ones, to me, are that a pluralistic view needs a consistent set of background assumptions and that a theory of social cognition involves the important and fundamental claims from the interactive turn. In what follows, I will frame social cognition as a phenomenon that entails several mechanisms spanning from deeply embodied, to representational and phenomenal. Throughout the course of this chapter, it will be clarified how 1-3E and PP can be merged to form the theoretical bedrock which 1-3sE will draw on. After detailing the basics of my proposal in chapter 7.1., I will then proceed to detail each level of embodiment and show how this emerging new picture may incorporate the desiderata and components of social cognition that have been worked out in Part I of this thesis.

7.1. The Basics of 1-3sE

This subchapter serves to make the reader acquainted with the goals of my proposal of 1-3sE. In two steps I will introduce the framework that will be described in more detail in the next chapters.

[1] Some content of this chapter has been published in Quadt, 2015.

(1) In chapter 7.1.1., the purpose and scope of 1-3sE is presented. I will explain that a phenomenon that is as manifold and diverse as social cognition needs a unifying view and scrutinize how this can be achieved by framing social understanding in a hierarchical framework.

(2) The next subchapter 7.1.2. serves to briefly summarize each level of embodiment and thereby give the reader a first idea of what to expect from the following more detailed depiction of 1-3sE.

7.1.1. Purpose and Scope

What is the exact purpose of 1-3sE? In Part I, I have argued that theoretical variety in the field of social cognition poses several challenges, and that current theoretical approaches do not meet the desiderata that have been worked out. Further I claimed that all of the existing theories make important and valid points about social cognition and that this leads to the task of finding a theoretical framework that puts them together in a coherent way. To suggest possible ways of doing so is the main goal of my proposal here. In other words, I aim to provide a framework which fulfills the desiderata, goals and challenges which have been described in Part I. I shall now detail the several ways in which this framework may be utilized in achieving this aim.

First, I suggest that it can serve as a scaffold to frame and describe the components of social cognition that have been worked out, as well as processes, mechanisms and experiences of social beings at different levels of analysis (d_3 – multi-level analysis; d_7 – comprehension). In this way, 1-3sE aims to depict social cognition as a diverse and profoundly dis-unified phenomenon that requires to be analyzed at a range of different levels. At the same time, the framework is supposed to yield a unifying perspective. This statement needs careful clarification, for my goal is not to present a 'grand unified theory' of social understanding which explains it all. Much rather, the unifying power of this theory consists of its diversity. *Prima facie*, this claim seems paradox, however, this paradox can be eased by explaining what exactly I mean by unification and diversity.

To find a unifying theory, here, first means to find a theoretical framework which operates on a set of non-contradictory background assumptions. This is, or so I have argued, a profound desideratum for any theory that puts together diverse elements.

1-3sE is unifying because it allows to pick a variety of social phenomena and analyze them with a set of conceptual tools that can be operationalized at different levels of description (d_4 – terminological consistency). In virtue of its multi-level, hierarchical structure, 1-3sE thus allows the analysis of a diversity of phenomena, while operating on a set of coherent background assumptions (d_6 –consistency).
This set rests on a merger of the original 1-3E framework and principles of PP and shall be applied to social cognition. By merging 1-3E with PP and applying this fusion to social cognition, I wish to present a theory that considers both low-level processes which are supposedly highly determined by an organism's actual physiology, as well as more abstract and less directly bodily influenced mechanisms (d_8 – integration). This is made possible by the hierarchical nature of the theory, which depicts an increasing abstraction and detachment from an individual's physiology and external signals as one goes up the hierarchy. I suggest that the combination of 1-3E and PP, applied to the study of social cognition, delivers a view that lives up to the high standards a theory of social cognition needs to fulfill. PP suggests that the functional and cortical hierarchy reflects a growing abstraction from the actual sensory input, as described in chapter 6.5.2. The same applies to social cognition, or so I will argue. While 1sE exhibits social processing that highly relies on actual sensory input and is profoundly shaped by the actual physical body, body representations described at level 2sE are not as much determined and constrained. This becomes obvious when considering the plasticity of the body model and the degree of abstraction needed to exploit one's own body model for understanding others. At the level of 3sE, we can observe not only an even higher degree of abstraction, but also the processing at larger spatiotemporal scales. In this sense, too, its hierarchical architecture gives 1-3sE its unifying power; by its very nature it assigns a fundamental role to the body, but then climbs up to more abstract processing.
Additionally, I hope that 1-3sE enables the analysis of a given social target phenomenon in more detail. The general idea is to take a social phenomenon, such as the experience of immediately understanding a person, and then scrutinize it at different levels of descriptions. For example, one would describe a specific social experience in phenomenological terms at the level of 3sE, and then find possible computational counterparts (2sE). In an additional step, it is possible to ask – for

example – whether the phenomenology changes depending on the morphological similarities between agents. Do I experience myself as being quicker to detect your emotions when your facial features are morphologically more similar to mine? Or does that difference merely constitute itself at lower levels and remains unavailable for conscious awareness? Embedding phenomena in this framework then not only yields new ways to ask questions, but also enables to hypothesize about relations between levels of embodiment. How does morphology change representational processing? And how do these changes relate to alterations in phenomenology? What will be needed for this endeavor to be successful is an interdisciplinary approach, that is, an approach which is able to integrate insights from both empirical sciences as well as theoretical considerations (d_2 – interdisciplinarity).

Recall that I presented two versions of the original framework of 1-3E; the systems and the hierarchical view (see chapter 5.1.). In the systems view, levels of embodiment are used to classify systems, while the hierarchical view can be recruited to analyze the behavior (for example) of one system at different levels of description. Of course, the framework I introduce here can be used in the same twofold manner. Fireflies who synchronize their blinking patterns, for example, may be described as 1sE systems, dynamically forming their pattern and exploiting their morphology for this joint behavior. However, I will focus on the hierarchical view in this thesis and merely touch on the systems view.

7.1.2. A Brief Description of Levels of Social Embodiment

After presenting these general ideas, let me now briefly describe how I aim to apply them to the study of social cognition and by doing so generate my own proposal of a framework, namely 1-3sE. The overall scope and purpose has already been discussed, what remains to be detailed now is what each level of social embodiment amounts to and how the merger of 1-3E and PP can be made fruitful for social understanding. At large, the three elements discussed in chapter 6.5.4. are adopted. These were, to briefly recapitulate, a refined notion of representationality as a gradually emerging phenomenon, the claim that levels of embodiment are used as heuristics to mark processing stages in the functional hierarchy, and the assumption that representations are grounded in sensorimotor processing. In the course of the following chapter, it will become more obvious how these elements relate to each level of social embodiment. I will now summarize what each level of social embodiment exhibits, in order to give the reader an idea of what this framework will look like, and also begin to suggest how the three elements will be implemented.

First-order Social Embodiment (1sE) describes low-level processes that ground higher-level ones. I will elaborate on the physical implementation of more abstract computations which enable social understanding. The questions to be asked and answered (at least tentatively) with regard to the issues that have been worked out in Part I of this thesis are: Which role does the body of an individual play for social understanding and in which way are bodily resources recruited to enable social cognition (a_7 – embodiment)? How should we conceive of the 'social brain' (d_5 – empirical plausibility; d_2 – interdisciplinarity)? Do interactive dynamics constitute social cognitive processes (a_8 – interaction)? Three central notions that shall begin to answer these questions will be introduced and discussed at the first level of embodiment: *embodied social inference (EmSI;, neural reuse, and interactive inference (InI;* a_1 – inference*)*. I distinguish two types of InI, namely *reproductive interactive inference* (*RInI*; a_6 – reproduction) and *complementary interactive inference* (*CInI*). These concepts will be elaborated on in the next chapter, but there are some underlying assumptions I shall name at this point.

In accordance with the claim that representations are grounded in sensorimotor processes, it will be claimed that bodily changes will lead to representational changes. This is important since I state that, on the one hand, bodily similarity appears to form a fundamental bridge between individuals (a_2 – similarity), and, on the other hand, differences in bodily constitution create an epistemic gap between agents. This in turn relates to higher levels of embodiment in the sense that representations which are used for higher order social cognitive processes are shaped by an individual's body.

Second-order Social Embodiment (2sE) refers to the computational level of description. The central question here is in which way the body model can be shared in order to make sense of other people. The notion of shared representations will be discussed in detail and several problems that arise are named; the *naked problem*, the *genuineness problem* and the *disambiguation problem*. In order to alleviate them, ideas from the PP framework will be recruited. Again, it will become clear that the body functions as a reference frame and hyperprior, thus yielding a meaningful self-other distinction which is necessary in order to make the sharing process not ending up in self-other confusion (a_3 – self-other distinction). This also entails to discuss the role of multisensory integration and, more specifically, the role of interoception for social cognition (a_9 – emotions).

Lastly, Third-order Social Embodiment (3sE) as the phenomenological level of analysis is discussed. Here I will scrutinize the range of experiences of social encounters (a_5 – experiential quality) and apply the concepts of transparency and opacity. In short, experiences of immediately understanding the other person will be described as transparent social states, while opaque states relate to the experience of explicitly constructing reasons for another person's behavior (c_4 – construction). The concepts of transparent and opaque social states will further be enriched by putting them into terms of predictive processing of sensorimotor contingencies (PPSMC, see chapter 6.5.1.). It will also be argued that opaque social states deserve their own level of embodiment, since they exhibit a rare and special phenomenon and draw on an additional representational process in the sense that the system phenomenally represents the underlying process itself. To do so, the level of 3sE+ will be introduced.

7.2. First-order Social Embodiment (1sE)

In this chapter, my goal is to describe the three following ways in which the lowest level of embodiment – 1sE – provides the basis for higher level social understanding:

(1) In an attempt to clarify the role of embodiment for navigating the social environment, the concept of *embodied social inference* (EmSI) will be introduced in chapter 7.2.1. The body will turn out to inhabit a twofold role for social understanding, namely as forming a bridge between individuals, but also creating an epistemic gap.

(2) In chapter 7.2.2., I will tackle two issues, namely the problem of finding an interdisciplinarily useful terminology and the question of whether and at which level social cognition is special or not. To do so, I will introduce the theory of neural reuse (Anderson, 2010) and apply it to research on the so-

called social brain. I claim that this will help to find ways to alleviate several problems that have been worked out in previous chapters in that it provides novel ways to tackle them.

(3) The last subchapter 7.2.3. deals with interaction dynamics and how they can serve to minimize prediction error in a fast and efficient manner. I situate social cognition within the framework of predictive processing. Specifically, I introduce the term *interactive inference* (InI) and discuss two types of it, namely *reproductive interactive inference* (RInI) and *complementary interactive inference* (CInI).

7.2.1. Embodied Social Inference (EmSI)

This subchapter deals with the role that the body plays for social interaction and cognition – both in an enabling and constraining way. Putting together the results from my investigation of Part I, I here argue that physical bodies create a fundamental bridge, but also create an epistemic gap between individuals.

As I have described in more detail in chapter 2.1.3., phenomenologists such as Merleau-Ponty argue that intercorporality is the *constitutive basis* for intersubjectivity, meaning that our bodies are the basis for the possibility to understand other people. In a deep sense, the body is seen as an 'inter-individual bridge' in the phenomenological tradition. In this chapter, I will pick up this idea and argue that our bodies are indeed profound for social cognition. Although I agree with the claim that the body plays a major role for understanding others, I reject the assumption that bodies create a *seamless* bridge and that thus there is no epistemic gap between individuals.[2] In general, I will argue that the body *determines* the degree to which individuals understand each other and in what kinds of interaction they can engage. This determination – and here is where my stance departs from the phenomenological one – goes in both a positive and negative direction. On the

[2] For the denial of such a gap, see, for example, Zahavi, 2011, pp. 550–555: "Rather the point is to recognize that expressive phenomena are already from the start soaked with mindedness. […] If phenomenological analysis tells us that the perceptually observed expressive phenomenon is already saturated with psychological meaning and that this is the explanandum, we should reconsider postulating mechanisms supposed to bridge a nonexistent gap."

one hand, our bodies enable our species-specific social abilities, in the way that the gross bodily structure determines the way human individuals are able to interact with and make sense of each other. Also, our bodies can be seen as grounding and forming the basis for motor and body representations which can be shared with others in the sense that their content partly overlaps and is suited to be used for both self- and other-related processing.[3] On the other hand, or so I will argue, they create an epistemic gap between individuals, thus constraining inter-individual action and understanding. To conceptualize these claims, I introduce the term of embodied social inference (EmSI). In accordance with a PP and FEP stance, I suggest that the notion of embodied inference can be applied to the social domain in the form of EmSI. The notion of EmSI will be laid out in more detail in the course of this subchapter and will help to describe the fundamental role of the body for social cognition.

What exactly do I mean by EmSI? To clarify the term, it will be helpful to recall the related notion of embodied inference. Embodied inference means that the thermodynamical laws of an agent's environment are 'folded into' her morphology; that her very physical body is built to keep her alive by resisting the second law of thermodynamics (see chapter 6.2.1.). In this sense, it can be said that the agent is a model of its world. This is related to the claim that the physiology of an organism determines the kind of mind it has, because the laws that are relevant for this specific phenotype will be modeled by its body.

In the same way, it can be said that the kind of body an organism has determines the kind of social interaction and understanding it is capable of. While a herring strives to stay in its large fish school to ensure its survival, cats aim for much smaller groups or may even survive on their own. The human body needs a caretaker for quite a long time during its childhood, not being able to sustain itself until a certain age. Further, while humans are able to use their speech-apparatus to communicate and interact, ants will have to rely on pheromones to send signals to each other. This can be seen as EmSI; an agent is – in virtue of its specific bodily structure – a

[3] Note how this creates an interesting transition point between 1sE and 2sE. This will be picked up in later sections of this chapter.

model of its social environment, their phenotype determines the kind of social abilities they possess.

Further, while there are very many individual differences, the gross anatomy and morphology of individual organisms of one species is rather similar. This similarity may provide a fundamental condition of possibility to recognize the other as 'one of us' and thus to understand them. The role of similarity is twofold; it not only determines how well we understand another person, but it also opens up the possibility that there needs to be a general similarity for social processing to begin with. The claim that a certain degree of similarity is needed in order to understand each other has been famously formulated by a number of researchers. For example, Meltzoff (e.g., 2005, 2007, 2013) states in his 'like me' hypothesis that the development of understanding others hinges upon the fact that the infant perceives the other as 'like me'. In fact, it is claimed

> that the core sense of similarity to others is not the culmination of social development, but the precondition for it. Without this initial felt connection to others, human social cognition would not take the distinctively human form that it does. (Meltzoff, 2013, p. 139)

Meltzoff's reasoning rests on the assumption that social cognition – especially in developmental terms – is enabled by matching visual to motor representations. The bedrock of his argument are his own and many follow-up studies on neonatal imitation (Meltzoff & Moore, 1997). Although having no visual information about one's own face, newborn babies appear to be able to imitate an adult's behavior, such as tongue protrusion (Meltzoff & Moore, 1977). It is thought that the visual information of the adult is 'matched' onto the proprioceptive information the newborn already acquired. This matching process then enables imitative behavior. The 'like me' hypothesis gains additional support when viewed from a PP perspective. In accordance with a simulation model, Friston and Frith (in press, p. 12) argue that "internal or generative models used to infer one's own behaviour can be deployed to infer the beliefs (e.g., intentions) of another – provided both parties have sufficiently similar generative models." In other words, similarity here is seen as a presupposition for mental state inference. Only when there is a sufficient similarity of models, there can also be a big enough overlap which allows the application of one's own models to understand the other's behavior.

In order to see how all of this relates to EmSI, though, we have to take a step back. For while the inference of beliefs and intentions is of course a rather high-level skill, the crucial point is to consider how these generative models are acquired and what they are based on. As described in chapter 6.1.2., generative models largely draw on prior experience. This experience is in many ways determined by our bodies. They are not only the spatial reference frame for our perspective, but our bodily abilities (and disabilities) also determine the richness of our motor repertoire. Clark (2016, p. 175) thus argues that "[…] our basic evolved structure (gross neuroanatomy, bodily morphology, etc.) may itself be regarded as a particularly concrete set of inbuilt (embodied) biases that form part of our overall 'model' of the world." In this sense, the body can be regarded as a central 'hyperprior' which largely determines what kinds of predictive models can be generated. In other words, it defines physical boundary conditions and constrains the space of computational possibilities.

If it is true that anatomical as well as morphological features are the basis for a system's generative models, and if it is true that these models can only be used for both self- and other-related processing if they are sufficiently similar, it follows that the bodily basis must be sufficiently similar, too. Put differently, if the bodily structure of individuals is grossly[4] different, their models may not be sufficiently similar, thus restricting interaction and understanding. The relation to EmSI should be clear by now; the phenotype of an individual must exhibit some degree of similarity in order to make it possible to recognize others as 'like me' and thus to enable the matching of one's own models to the other's.

However, the claim that bodies and bodily abilities are the basis for generative models not only implies that those will be somewhat similar. Since bodies are also considerably distinct in many other respects, this idea can also run in to the other direction. This leads me to the second part of my claim, namely that our bodies form an epistemic gap between individuals. While what I have presented so far rather speaks for the fact that EmSI determines social behavior and abilities in a

[4] Consider that I claimed earlier that there must be a *gross* difference in bodily structure in order not to be able to understand each other at all. Of course, a father will still be able to somehow understand and interact with his newborn child, although their body differ considerably in size, strength, etc. Thus, this similarity basis can only be very rough.

positive way and thus for the claim that our bodies form an inter-individual bridge, they also create a fundamental gap between individuals and thus constrain our social abilities. Just as the gross structural similarity of our bodies enables the generation of sufficiently similar models, there are several factors which lead to individual differences that – or so I will argue – constrain social understanding.

While (most) human bodies are quite similar with respect to, for example, configuration of body parts and relative limb size, finer physiological features such as overall size, weight and skin color can vary considerably. These variable features, though, are likely to form an important basis for the generation of predictive models. Again, the basic assumption is that *physiological* differences will result in *representational* differences.[5] While generative models that are built on a bodily basis are thus sufficiently similar to be partially shared, they will most likely never be identical.

Further, one prominent position is that bodies form the spatiotemporal reference frame of our experience. Most of these views, though, focus on the *phenomenal* perspective and the role our bodies play here. This does not mean, however, that bodily structures do not determine low-level processing, too. Whether or not an individual is conscious and whether or not sensory signals reach the level of awareness, the body is the medium through which an organism interacts with its environment and gathers its knowledge. What I mean here by spatiotemporal reference frame, thus, is the very basic, simple fact that our bodies determine the when and where of an organism's environmental interactions. If my body is sleeping in Germany, I cannot witness the earthquake in California at the same time. This seems trivial, but it basically means that bodies determine the range of experiences one can have. Metzinger (2004, pp. 160–161) picks up this point and formulates it as the 'single-embodiment constraint':

> Trivially, the causal interaction domain of physical beings is usually centered as well, because the sensors and effectors of such beings are usually centered within a certain region of physical space and are of limited reach. […] This functional constraint is so general and obvious that it is frequently ignored: in human beings, and in all conscious systems we currently know, sensory and motor systems are physically integrated within the body of a single organism. This singular

[5] These representational differences are of major importance for 2sE and their effects will be described in the following chapter.

> "embodiment constraint" closely locates all our sensors and effectors in a very small region of physical space, simultaneously establishing dense causal coupling.

In making this statement, Metzinger clarifies that the behavioral space of an individual is limited and constrained by its body. The range of possible behavior and experiences shape our cognitive processing, an effect whose pervasiveness becomes clear when viewed through the lens of PP. Recall that PP depicts the neural and cognitive architecture as immensely flexible and ever-changing. If precision-weighting admits, any sensory signal can change predictions at any level of the processing hierarchy.

Related to this thought is also the 'asymmetry of access' claim by Lipps and Scheler which has been described in chapter 2.1.1. To briefly repeat, both philosophers argue that we do not have the same kind of access and information about others than about ourselves. While other-related information is exteroceptive, mostly visual input, self-related information includes proprioception and interoception. The latter thus seems to be a richer, or at least a very different ground to build generative models on. All these aspects, or so I suggest, lead to a different body model, given that this model is built upon individual input. In chapter 7.3.1., it will be argued that the body model can be used for social cognition and that these individual differences will play an important role here. It thus seems obvious that bodies *ground* higher-level representational processing and provide the basis for the body model.

So far I have argued that EmSI leads to a twofold role of the body as forming both a bridge and a gap between individuals. This claim is of importance for another fundamental theme in research on social cognition, namely the distinction between self and other. When it comes to this topic, it appears to be common sense that such a distinction is needed to enable inter-individual understanding. Where this agreement usually ends, however, is on the question of when and at which level a self-other distinction arises. Applying this point to the theoretical framework of this dissertation, one can ask at which level of the 1-3sE hierarchy a distinction between self and other arises. Is it *sufficient* to be an embodied organism with a somewhat unique morphology to distinguish self from other? The original 1-3E framework seems to suggest that social understanding and a self-other distinction only arise at the level of 2E and requires a unified model of oneself and one's body.

However, if EmSI is indeed at play, could we say that very rudimentary forms of a meaningful distinction between self and other is enabled at an even lower level of processing? Could it be that “[t]he mere fact of embodiment could be the first level at which this distinction between oneself and other subjects or objects takes place”, a possibility that Tsakiris and Fotopoulou (2008, p. 1318) propose? The question here seems to boil down to whether or not a body model that rests on higher-level integration processes, as described at the level of 2E, is needed in order to sufficiently demarcate *this* body from other bodies.

Given that our bodies are unique and individual in the sense described in this section, they yield sensory inputs which are not very likely to originate from any different source than *this* body. If PP is right in stating that probability distributions are pervasive throughout the processing hierarchy down until the lowest level it appears plausible that some sensory input is ‘more likely to be me’ than another even at the lowest processing stage. The combination of EmSI and IPP may provide the necessary equipment to support such a claim. Embodied inference describes the fact that the phenotype of a species incorporates the laws of its environment in order to maintain homeostasis. Maintaining homeostasis, however, is enabled by IPP, as described in chapter 6.4.2. This strongly suggests that the system ‘knows’ – in a very basic way – its boundaries; that an organism can differentiate between what needs to be kept alive and which force hinders it. Additionally, IPP indicates that while (most) interoceptive regulatory processes keep *this* body alive, they do not the same for other bodies. Of course, all of this can be equated with proper self-identification and a full-fledged self-other distinction. However, all of the above hints to the cautious stance that a crude distinction between self and other already takes place at the level of 1sE.

In this subchapter, I have argued for a determining role of the body for social cognition and in this way integrated component c_7 – embodiment. I further attributed a fundamental role to similarity (c_2) in the sense that a basic similarity is needed to establish a specific social relation between individuals. Also, physiological similarity has been hypothesized to end up in representational similarity, again providing a basis for social understanding. In what follows, I will turn to the role of the so-called social brain.

7.2.2. The 'Social Brain'

In this subchapter, I will elaborate on the notion of the 'social brain' as the neural implementation of social cognitive skills.[6] Several conceptual issues were distilled in Part I of this thesis that shall now be tackled. First, one rather general problem was that of finding an appropriate terminology which is approachable interdisciplinarily and thus helps to bridge gaps between research groups. I aim to suggest ways of how to alleviate these problems and thereby to fulfill desiderata d_2 – interdisciplinarity and d_4 – terminological consistency.

Second, recall that the social brain hypothesis (SBH, see chapter 3.2.1.) in one of its original formulations refers to the claim that there are anatomically fixed modules in the brain which support social processing (Gazzaniga, 1985). Although more current views tend not to subscribe to this strong modularity anymore, picking out regions or sub-regions and assigning a specific role to them still seems to be commonplace. The underlying conceptual problem is that these views provoke the question of how special social cognition is compared to more general cognition – and whether this specificity is already to be found at the neural level. I will argue that PP as one very promising current theory predicts a much more complex and interconnected architecture of the brain than most post-phrenological views. The hierarchical structure that is fostered by PP would much rather point to a neural organization that is both functionally and organizationally more flexible. The ever-present message passing from both the top down and the bottom up requires a massively interconnected structure of neural passaging. To further substantiate the take on PP, I will introduce Anderson's (2010) theory of neural reuse and his massive redeployment hypothesis. Anderson's view not only offers a way to start to answer whether social cognition is special compared to more general cognition (d_1 – specificity). It also yields a set of new, more specific terms to conceptualize the implementational and functional level of description. In addition, I will argue that it integrates core assumptions of PP and thus yields a smooth transition to the next level of analysis. In drawing on PP and neural reuse, another goal of mine is

[6] As this is a mainly philosophical work, I certainly do not aim to offer a full-fledged theory of neural architecture and implementation, but rather aim to address the conceptual problems and desiderata than have been worked out in chapter 3.

to fulfill desideratum d_5 – empirical plausibility. In more detail, instead of basing my take of the 'social brain' on post-phrenological views, I aim to include the claims of novel frameworks.

Let me start by briefly presenting the theory of neural reuse and the massive redeployment hypothesis.[7] In general, Anderson's goal is to present a theory on the functional architecture of the brain as an alternative to modular views of brain function. Much recent research has hinted at the functional diversity of brain regions and thus motivates theories that reject anatomical and strong functional modularity. In this manner – instead of assuming that functions are found in highly specialized anatomical modules – neural reuse states that

> […] circuits can continue to acquire new uses after an initial or original function is established; the acquisition of new uses need not involve unusual circumstances such as injury or loss of established function; and the acquisition of a new use need not involve (much) local change to circuit structure (e.g., it might involve only the establishment of functional connections to new neural partners). (Anderson, 2010, p. 245)

In other words, neural circuits are recruited and re-recruited for diverse functions, and while they have an *original* function, they can acquire *new* and *different* functions when connecting to other circuits.[8] This view (which has been described in brevity here, for a much more detailed depiction, see Anderson, 2010, 2014, 2015) has several important implications.

The principle of reuse implies that if there is not one specialized region for one task (e.g., TPJ as the ToM-module), then a brain region should correlate with a variety

[7] Please note that the evidence for neural reuse as described by Anderson is still sparse, partly due to the fact that this is a fairly new approach. For a more detailed discussion, see Anderson, 2010, 2015.

[8] The principle of neural reuse applies to both the ontogenetic and the evolutionary development of the brain. From an evolutionary perspective, the massive redeployment hypothesis states that the principle of reuse is realized in the re-recruitment of already existing and functionally assigned brain matter during the acquisition of new skills. A strong modular view could hardly explain how ever more cognitive skills could have evolved without adding new specialized regions in form of a considerable amount of additional brain tissue at the same time. Note, though, that this does not mean that new brain tissue is never added – the crucial point is that already existing structures will be used whenever possible. For reuse is much more parsimonious and less 'costly' than the growing of ever bigger brains. The same applies to ontogenetic development. As individuals learn more and more skills, their brain develops with them. Instead of growing more brain tissue for every new skill to form a new module, neural reuse instead suggests that already existing areas are redeployed.

of tasks. In order to investigate this claim, Anderson and colleagues (2013) looked at the activation patterns of 78 standard anatomical regions and calculated whether and how often they are active in 1138 experimental tasks that were used in 11 different Brain Map task domains.[9] The results show that on average, a brain region is active during 95 tasks out of 9 of the task domains. Thus it is concluded that "local neural structures are *not* highly selective and typically contribute to multiple tasks across domain boundaries." (Anderson, 2014, p. 10)

According to these findings, one may object, is there no specialization in the brain *at all*? At this point, it will be helpful to clarify that although strong modular views are rejected, the theory of neural reuse does not deny that brain regions have intrinsic functional *dispositions*. In order to conceptualize this claim, Anderson and colleagues (2015, p. 4) introduce the term of a "functional fingerprint", which is based upon the results of a follow-up study of the meta-analysis mentioned before. In this study, the likelihood of brain regions to be activated in a task was calculated. This likelihood is referred to as the "functional fingerprint" of a brain region, where the term is defined as "the likelihood that an active region is active during, or activated by, a given type of task or stimulus, and thus offers a way to capture the different functional biases or underlying causal dispositions of individual regions." (ibid.)[10] The notion of a functional fingerprint thus expresses that neural circuits are 'specialized', in the weak sense that their functional disposition makes them more likely to be recruited for some but not other tasks. This term is one of many that Anderson introduces and which helps to build a terminology that yields a conceptual dissociation from views that foster strong specialization.

Another useful concept which is related to functional fingerprints is that of "functional differentiation" (Anderson, 2014, p. 16). The term in introduced in order to find an alternative expression for 'specialization', since this is seen as too radical. Rather, "local neural assemblies persist in supporting many tasks across ostensibly quite different task categories […] and thus retain a complex response profile." (ibid., p. 52) In other words, it is not rejected that brain regions are

[9] These were: action execution, action observation, action inhibition, attention, audition, vision, emotion, language semantics, reasoning, explicit memory, working memory.

[10] Furthermore, this method shows the degree to which a region is a 'specialist', since it indicates the number of tasks during which it is active.

completely undefined. They are, though, functionally differentiated in virtue of implementing different response profiles.

But still one question remains, viz. how is it possible that ever more skills are implemented, ontogenetically as well as phylogenetically, without producing ever more brain matter? In other words, how is functional differentiation achieved? The solution lies, according to Anderson, in the highly interactive manner in which local neural assemblies connect with each other. To express this conceptually, he refines functional differentiation even further and introduces the term of "interactive differentiation" (Anderson, 2014, p. 58). This means that functions arise in and are differentiated in virtue of *interactions* between regions, thus "[…] function depends much more on the *interactions between* parts than on the *actions of* parts […]" (ibid., p. 40). In Anderson's (ibid., pp. 52–53) own words:

> […] local neural assemblies will come to have particular, distinctive response profiles, as determined by a combination of intrinsic local cortical biases and extrinsic factors including experience and the influence of functional interactions with other regions of the brain. A region's response profile will certainly reflect its underlying functional capacities and determine its role(s) it can play in various functional coalitions. But although this might therefore be considered a kind of functional selectivity, it is quite different from the notion that brain regions will come to specialize in such tasks as "face perception" or "mind reading" and furthermore suggests a developmental pathway relatively unconstrained by (or simply in some sense insensitive to) the traditional categories of cognitive psychology.

By rejecting that there are brain regions that are specialized for isolated tasks, and by emphasizing the vast possibility of influences on the implementation of functions, Anderson clearly sides with views that believe brain function to be widely distributed in a complex network-manner.

This has implications for the research on social cognition. It gives hints for how to incorporate the desideratum of specificity, which I will turn to now. Further, the terminology that is used by Anderson can be applied to social cognition and thus serves as a starting point to a common terminological ground between disciplines. Concerning the question of whether social cognition is special or not, neural reuse would clearly deny – as already stated in the quotation above – that there are brain regions which are *merely* specialized to process social stimuli (for it is denied that there is *any* kind of exclusive regional specialization). However, this does not imply that social cognition is not 'special' compared to general cognition. What is on offer here, it seems to me, is not a rejection of the claim that social cognition is distinct. Much rather, it offers a more fine-grained and cautious approach to the 'social brain' and also provides a more differentiated terminology.

One possible way for future research, for example, could be to look for 'social functional fingerprints' (SFF). Stanley and Adolphs' (2013) work can be interpreted as an attempt to do so. They assume that current views on the 'social brain' – such as those presented in chapter 3.2. – can be biased by already existing theoretical and conceptual predispositions in social psychology. In order to avoid these biases, the authors used a data-driven approach to find 'social' networks in the brain. In a first step, reverse-inference maps were derived in order to find the likelihood that a study used the term "social" depending on whether or not a specific type of activation was present. Subsequently, 200 independently identified topic maps (topics included emotion, social games and interaction, fear and arousal, consciousness and awareness) were analyzed. The authors sorted out maps that (a) were based on more than 30 studies, and which either (b) covered more than 50% of the first map (the "social" term map), or (c) was covered more than 50% by the "social" term map. Interestingly, the results fit into what neural reuse would predict: While regions like PFC and the amygdala seem to be involved in a broad range of social tasks (and may thus play a more general functional role), other areas such as the precuneus seem more 'selective'. In accordance with these datamining results, the authors conclude that

> what has emerged from the corpus of social neuroscience research is not a single, but several, neural systems for processing social information. Correspondingly, there has been a shift from focusing on the function of structures in isolation […] to understanding circuits and systems, with increasing attention to connectivity […]." (ibid., p. 821)

In other words, future research will most likely shift from looking for isolated 'social' areas to the search of networks that are widely scattered and more flexible. Note that the method that was used here can be seen as a stepping stone between more classic and entirely novel views. For one possibility is to find mappings between results from these neuroscientific meta-analyses that yield new insights and preexisting concepts in social psychology (cf. ibid., pp. 821–822). Data-driven approaches such as the one just presented can function to gather new results, but also to compare them to already existing ones. Thus, instead of finding a substitute for current methods, it seems to me, the authors offer a bridge that may facilitate the transition to a new view of the social brain. This new view emerges when more focus is put on finding SFFs, i.e., local neural circuits which have the functional disposition for processing social stimuli (but may well inhabit other functions, too). The concept of SFF is especially useful, as I see it, because it emphasizes that some neural assemblies are particularly well-suited for social stimuli, thus hinting to an answer to the question of whether or not social cognition is 'special' in any sense. It thusly suggests that function is an appropriate criterion for individuating brain networks and supports the claim that what makes social cognition 'special' is its function, i.e., to process social stimuli.

Interestingly, this perspective may make interdisciplinary communication easier. One problem, as has been described in more detail in chapter 3.3.2. that arises in interdisciplinary research is the need for a common language which presupposes a mapping of concepts of one discipline to those of another. Stanley and Adolphs (2013, p. 822) propose that

> [a] large part of this tension stems from the belief among some social scientists that the processes responsible for understanding both human and animal social behavior are very complex, are very context-dependent, and draw on many factors, including ones outside the brain—as such making these processes ill suited to neuroscientific study.

The problem, according to the authors, thus lies in finding the 'appropriate' level of description for social cognition and finding bridging principles between levels. Looking for SFFs may offer an answer here. For an SFF is a neural network defined by neuroscientific study of the brain.

An interesting point made in the quotation by Stanley and Adolphs above is that context-dependence leaves social behavior ill-suited for neuroscientific research. However, when PP and neural reuse are applied, context enters the picture rather naturally, and in a way which allows contextual influence both at the personal and neural level. Social psychologists see 'context' as a situational factor, meaning that the social environment influences an individual's behavior and that this kind of influence can hardly be tracked down to the neural level. In a multi-level, hierarchical model such as PP, though, exactly this may now be possible. Contextual cues, as we have seen, influence processing not only at higher levels, but all the way down the hierarchy. Different contexts elicit different priors which in turn trigger different predictions:

> Basic context effects, within the PP framework, flow inevitably from the use of higher level probabilistic expectations to guide and nuance lower level response. This guidance involves expectations concerning the most likely patterns of unfolding activity at the level below. But such expectations […] are intertwined with context-based assessments of the reliability and salience of different aspects of the sensory information itself. (Clark, 2016, p. 146)

In PP, thus, contextual effects are central to neural processing and remain a pervasive influence factor at multiple scales. This will become of major importance when it comes to the level of 2sE and possible bridges between 1sE and 2sE.

In sum, whether or not neural reuse will prove correct and whether it will indeed be useful for research on social cognition is still up to debate. However, as I suggested, it seems to be a promising approach which yields numerous opportunities for conceptual clarification and improvement.

7.2.3. Interactive Inference (InI)

In chapter 2.3.5., I have casted doubts on the empirical and conceptual validity of the claim that interaction constitutes social cognition. However, the role of interaction for social cognition should of course not be underestimated. Opponents of representation-based mindreading accounts have claimed that one reason why it is fundamentally false to neglect interaction in any theory of social understanding is that most social skills are developed during interactions. Even throughout adult life, interaction remains the primary context in which social cognition takes place. In this chapter, I offer a different approach to emphasizing the importance of interaction. Instead of relating its importance to its quantitative number of occurrences, I claim that interaction is a highly efficient means for understanding other minds when viewed from the PP stance that I adopt here. Further, I will show how this perspective enables us to integrate the primacy of interaction, and thus to include interaction as a central component of social cognition, while not throwing away the concept of representations and inner models.

An important difference between phenactivist and mindreading accounts is that their perspectives have often investigated social processing at different timescales. While TT and ST have focused on slower, higher-level processes, phenactivism seems to have focused on interaction, whose highly dynamic unfolding may happen at fast and rather low levels. Of course, older versions of ST and TT and a rather classic cognitivist view on social understanding would not be able to cover such a complex and dynamic process. However, it appears that modern theories such as PP do have the repertoire to integrate interaction as a process that relies on representations – viz., action-oriented representations.

The trick is to acknowledge that the task of predictive models (i.e., representations) is to find the most efficient, least costly route to success. This is what Clark (2016, p. 244) refers to when he talks about the "productive laziness" of the brain; whenever the body or the world can be recruited to do a job, there is no need to compute complex inner models instead. Precision-weighting determines whether low-level modalities or high-level modelling will 'be in charge' to solve the task at hand – depending on how efficient the strategy is estimated to be. This strategy will more often than not involve the engagement of brain-external structures:

> The task of the generative model […] is to capture the simplest approximation that will support the actions required to do the job – this means taking into account whatever work can be done by a creature's morphology, physical actions, and socio-technological surroundings. […] There is thus no conflict with work that stresses biological frugality, satisficing, or the ubiquity of simple but adequate solutions that make the most of brain, body, and world. (ibid., p. 291)

By making this statement, Clark endorses a central aspect of phenactive theories, namely the role of brain-external structures for an agent's navigation. Active inference takes center stage in this interpretation of PP, in virtue of the fact that the function of predictive models is to distribute the cognitive workload and recruit embodied action whenever possible. While I generally agree with his view, I wish to add that interaction can play the same role here for social cognition – namely the role of gathering information about the social environment and thus actively sculpting not only one's external, but also internal environment. Thus I claim that while *active inference* is central for general cognition, *interactive inference*, as I will call the process, is just as central for social cognition.

What exactly does 'interactive inference' mean? Active inference has been described in chapter 6.3.1. to minimize prediction error in several ways, namely by actively changing an agent's inner and outer environment so to fulfill exteroceptive, proprioceptive and interoceptive predictions, and the disambiguation between competing predictive models. In a similar way, interactive inference can be described to minimize prediction error while navigating the social environment. Instead of changing one's model about the other person in order to understand her (perceptual inference), interactive inference serves to actively sample proof for predictions or to cancel out possible models about the other person. This happens in an interactive context, meaning that two (or more) individuals actively exchange information.[11] Basically, interactive inference can take place in any kind of encounter between two embodied agents.

First, consider what I will call *reproductive* interactive inference (RInI). This kind of interactive inference occurs during phenomena such as synchronization, entrainment, automatic imitation or emotional contagion which occur automatically and involuntarily; even when people are explicitly asked to suppress these tendencies. There are, for example, many studies which show that individuals cannot resist but synchronize their movements with the other person. This has been shown for several motor acts, such as finger tapping (Oullier et al., 2008), rocking in rocking chairs (Richardson et al., 2007) and body posture (Lafrance & Broadbent, 1976).

Chartrand and Lakin (2013) provide a comprehensive review on these effects and summarize them under the notion of 'the Chameleon effect': "[…] much like chameleons change their color to blend into their surrounding environment, humans alter their behavior to blend into their social environment." (ibid., p. 288) In a vast number of studies, it has been shown that mimicry and synchronization are accompanied by many facilitating factors, but in turn also facilitate social interaction. For example, individuals are more likely to mimic another person when there are prior 'pro-social' factors, such as in-group effects and prior rapport. Individuals with similar opinions and high empathy rates are more prone to mimicry and synchronization. Although there are also inhibitors of mimicry such as the wish to disaffiliate the other person, the authors conclude that unconscious mimicry and

[11] One rather obvious, but complicated question arises at this point: What counts as an 'interactive context' and what does not? While it seems clear that real-life face-to-face interactions count as such, the line gets more blurred when considering communication through text messages or even physical mail. I would suggest that all these count as interaction, but happen at different time-scales. However, for the sake of simplicity, I will here consider the obvious cases and leave the issue to future conceptual research.

synchronization seems to be a default for social interactions and occurs even when individuals face other tasks (cf. Chartrand & Lakin, 2013, p. 290). Further, individuals that were told to keep still and suppress their tendency to replicate the other person's behavior perform worse at emotion detection tasks.

How does all this relate to interactive inference? To see this, consider that synchronization, mimicry, or imitation are ways to make oneself and the other person more similar. Reproducing the other person's bodily states could be a highly efficient and quick way to get 'first-hand information' on their current condition. Instead of going through the hassle of generating brand new models, it may be more efficient to actually 'put oneself into' the other's bodily state. Since interactions unfold in a fast manner, the quickest way to gather information about the other person, to tune into their mood or intentions and thus to be able to react appropriately may be to simply replicate their body with my own.

In order to get a sense of the other person, predictions about their current state are corrected in virtue of error signals. When replicating the other's bodily state, these error signals should be more reliable, since they come not only from one exteroceptive (i.e., visual) source, but also from an internal source (e.g., proprioceptive prediction error). Therefore, during RInI the bodily state (e.g., posture, movements) of another person is mimicked in order to supplement exteroceptive information about them with interoceptive and proprioceptive information. Mimicry, synchronization and automatic imitation are instances of RInI that function to make predictions about the other more precise by increasing the number of signal sources that yield relevant information.

Furthermore, consider the following study conducted by Ainley and colleagues (2014) to link interoceptive awareness and the tendency to automatically imitate. They found that – contrary to their initial prediction – participants who scored higher for interoceptive awareness had a greater tendency to imitate. In other words, the more one is aware of her interoceptive processing (in this study measured with the so-called 'heartbeat perception task'), the less she is able to inhibit automatic imitation. One possible (although rather speculative) interpretation of these results is that people with higher interoceptive awareness set the gain on interoceptive prediction errors higher. Ainley and colleagues (ibid., p. 26) hypothesize that

> [g]iven that interoceptive awareness affects perception of the body, it is also likely to modulate action representations. It has recently been indicated that in order to avoid mirroring another person's actions it is essential to *reduce* the precision of proprioceptive prediction error (Friston, Mattout & Kilner, 2011). If people with high interoceptive awareness have initially precise proprioceptive prediction errors then their tendency to imitate others may be accounted for.

Put differently, in order to inhibit imitation and not to replicate the other's movement, gain on descending prediction error must be set low. Thus, weighting the precision of prediction errors high may result in the tendency to automatically imitate the other person. If this is correct, the scenario of automatic imitation could be the following. First, contextual cues yield information that the current incoming signals originate from another person and shared models about sensory consequences, which could be proprioceptive, interoceptive, or exteroceptive, are recruited.[12] The next step is crucial. Depending on whether the gain on prediction error is set high or low, the observed state of the other person is replicated or not. As described above, highly precise errors would result in a replication of the other's state, while low-weighted prediction errors would result in the inhibition of automatic imitation. This may not only be the case in motor imitation. Phenomena such as emotional contagion or the queasiness one feels when observing someone eating something truly disgusting could be cases in which gain on interoceptive prediction error is set high. This would lead to the replication of the other's interoceptive state and thus trigger 'shared bodily experiences'.

Automatically replicating the states of the other person is of course not the only process happening during interactions. It will often be necessary to perform complementary actions, which I will call *complementary* interactive inference (CInI). This second case of InI entails changing one's external or internal environment in response to the other person and in order to achieve a change in the other. CInI has several functions.

It firstly functions to regulate the other person's current bodily or emotional state. By changing one's own posture, movement, or gestures, or by altering one's interoceptive state, a responsive alteration in the other is triggered. An intriguing example can be found in so-called 'kangaroo care'. During kangaroo care, mothers

[12] For a more detailed description on the role of context for shared generative models, see chapter 7.3.4.

hold their (mostly prematurely born) infants in an upright position close to their body between their breasts and underneath their clothing. It has been found that this has many positive effects on both mother and baby. Most interestingly for the matter here are the physiological effects; mothers regulate their body temperature according to their infants needs and thereby also enhance self-regulation of the child. When the child has a fever, mothers lower their body temperature so to provide cooling for their infant. Further, the irregular heartbeat of a baby can be counteracted and becomes steadier when their ear is placed on their mother's chest and they hear the mother's steady heartbeat (Ludington-Hoe et al., 2006; Nyqvist et al., 2010).

Secondly, CInI can serve to evoke a behavioral response of the other, which then provides additional input for disambiguating the social stimulus. Facial expressions, gestures and other movements are used as signals to the other person that one is uncertain and needs further information. Shrugging my shoulders or raising my eyebrows, for example, may signal you that I did not understand what you were saying, meaning that additional information is needed. If the interaction is successful, this will cause you to elaborate on your stance.

A third function of CInI could be to make oneself more predictable and thereby smoothening social understanding, coordination, and joint action. Vesper and colleagues (2010, p. 999) introduced the term 'coordination smoothers' to describe modulatory behavioral processes that make coordination in joint actions easier:

> One way to facilitate coordination is for an agent to modify her own behavior in such a way as to make it easier for others to predict upcoming actions, for example by exaggerating her movements or by reducing the variability of her actions.

Accordingly, several studies show that people adjust their movement trajectories and pace or use signaling and communicative actions to increase predictability. Piano players in a duet, for example, exaggerate their finger movements or increase speed in order to decrease variability (Keller, Knoblich & Repp, 2007).

What are the mechanisms that underlie CInI? A hint can be found in Vesper and colleague's (2010) work. The authors claim that prediction and motor simulation are key processes which enable the execution of complementary actions. Simulations are thought to enhance timing and anticipation of sensory consequences, thus being especially important for joint actions. This can naturally be operationalized under a PP perspective, because it is assumed that predictive models represent sensory consequences of actions in a counterfactual manner (Seth, 2014). If it is additionally assumed that these models can be shared in the sense that they can be used for both self- and other-related processing, they can also be exploited to compute the consequences for one's own and the other's sensorimotor trajectory.

There seems to be, again, a role for similarity in these processes. For joint action is enhanced when the timing patterns of both agents are predictable. This predictability, in turn, depends on the similarity of agents and their motor experience. Several studies showed that mirror neuron activity increases when observing actions that belong to one's own motor repertoire (Calvo-Merino et al., 2004). The mirror neuron system is therefore involved in both replicative action processing and the preparation of complementary actions. According to Pezzulo and colleagues (2011, p. 612), "this suggests that the brain can encode actions executed by others in an interaction-oriented way, and more broadly that action-perception mappings could be quite flexible and task-dependent."

Taken together it can thus be hypothesized that shared predictive models are not only useful for replicative, but also complementary interactive inference. Interaction is here used to solve problems with the other person, in virtue of making oneself more predictable, and using one's body to signal what is needed from the other. We can now say that interactions involve different embodied dynamics, which means that the change of one's own state will contribute to the flow of the interaction and promote understanding. Other examples than the ones named above are changes in body posture and tone of voice, or social touch. All of these bodily signals can be used to trigger a reaction in the other person which will help to disambiguate the meaning of the incoming signals. Using our bodies in interactions thus may be a very efficient way for navigating our social environment. A further advantage of InI is that it is a quick and non-costly way to make sense of other people or to probe them, it enables a range of different sensory signals from different modalities to improve the probability of predictions. Additionally, it provides information that is genuinely from the other which will become important in the next chapter, when the notion of shared body representations is discussed. The present proposal underscores the importance of a_8 – interaction and a_4 –

reproduction in social cognition. It provides two concepts of mechanisms by which interacting partners can influence each other's models to aid in this interaction. The framework of PP provides the fast and dynamic mechanisms necessary to explain online social interaction.

7.3. Second-Order Social Embodiment (2sE)

2E, as described by Metzinger in the original framework, exhibits the representational and computational level of description. The kinds of representations we find at this level are more abstract than that of 1E/1sE, and also there is a higher degree of integration of sensory input from different modalities. This is, as I read Metzinger, the basis for the unifying body model which enables, for example, motor control. It is also the basis for several phenomenal experiences, such as body ownership and phenomenal selfhood. It has been claimed that the integration of multisensory information lies at the heart of these phenomena. Metzinger adds to the discussion that at the level of 2E, one can describe the representational body model which brings forth these experiences. These ideas serve as a basis for the following discussion points in this chapter.

(1) In chapter 7.3.1., I elaborate on the general idea of a body model and how it can be exploited for social cognition in virtue of containing shared representations.

(2) There are, however, some issues when applying 2E to social cognition. Three of them will be described in chapter 7.3.2. First, the claim that parts of the body model can be shared is somewhat problematic. In brief, if there is shared representational content that overlaps between agents, then how do we distinguish between self and other? This is what has been called the 'naked problem'. The second problem I will coin as the 'genuineness problem'. It refers to the possibility of merely understanding oneself instead of the other person if too much weight is put on shared representations. The third problem, the 'disambiguation problem' encompasses the more general issue of how incoming stimuli of other people can be disambiguated and thus used to infer their intentions.

(3) In section 7.3.3., the role of context in hierarchical predictive processing will be described in more detail and it will be scrutinized how this can alleviate both the naked and disambiguation problem.

(4) Chapter 7.3.4. deals with the question of whether others are mere versions of ourselves, an issue that occurs when the existence of shared representations is accepted. In accordance with the claim that our bodies serve as central priors for processing and thus yield a 'natural' self-other distinction, the answer to this question will be negative. Further, in virtue of the possibility to update predictive models with genuine other-information, the genuineness problem loses its sting.

(5) The last subchapter 7.3.5. concerns the role of interactive inference for disambiguating incoming stimuli. Once again, interaction will be displayed as a fruitful means to gather information about the other person.

In the following, I argue that the multisensory body model is not only crucial for self-related processing, but can also be exploited for social cognition. In more detail, I claim that it provides means for both sharing representations as well as distinguishing between self and other. In this sense, the body model functions as a bridge and a gap between individuals – a rationale I already laid out at the level of 1sE. This also creates a smooth transition between 1sE and 2sE.

7.3.1. The Shared Body Model

The general idea in this chapter is that parts of the body model can be recruited for social cognition and that the body model can thus become a *shared* body model. In order to make sense of this idea, let me briefly elaborate on the general idea of a body model. Of course, there are many different definitions of the term out there. One feature that is repeatedly found in leading theories on selfhood and its underlying representational processes, though, is that the body model contains *integrated multisensory information* from different modalities.

The idea that multisensory integration underlies the body model and thus underlies phenomenal selfhood can be found in the original 1-3E framework. Metzinger (2014a, p. 273) describes it as a "grounded, predictive body model that continuously filters data in accordance with geometrical, kinematic and dynamic boundary conditions.” In other words, features of one’s own body are represented and according to those, sensory input is processed in a predictive manner. Together with Blanke (2009), he argues that the integration of local bodily features leads to a global representation of the body, which then grounds phenomenal selfhood.

In chapter 6.4.2., I showed how Seth describes IPP to underlie a sense of self. In his account, the process of integrating interoceptive, exteroceptive and proprioceptive information are deemed the grounds for a sense of self. Another account already mentioned and even more important for the application to social cognition is Apps and Tskaris’ (2014) theory of the ‘free-energy self’. In an attempt to put multisensory integration into PP terms, they state the following. First, one has to accept the general claim that representations of the properties of one’s body are probabilistic, which is one core statement of PP. Further, the idea is that “one’s own body is the one which has the highest probability of being “me” as other objects are probabilistically less likely to evoke the same sensory inputs.” (ibid., p. 88) The crucial process here is the minimization of prediction error at all levels of the hierarchy, i.e., predictions about low-level, uni-modal input need to be tested and be congruent with prior, multimodal beliefs about one’s own body. Recognizing oneself in the mirror, for example, is thus possible when visual predictions are in line with other body-related predictions, for example, predictions about proprioceptive incoming signals when I move my hand. More generally speaking, multimodal representations of oneself, which are created in the process of integrating multisensory information, are crucial for a sense of self (cf. ibid., p. 86).

After this short detour, let me start with the claim that the body model can serve as a basis for sharing representations and thus to enable social cognitive processing. The general notion of shared representations is that there is an overlap in representational content which contains information that can be used for both self- and other-related processing. Put differently, some self-related representations (such as motor or body representations) can be applied to the other person, hence shared. In a similar manner, Metzinger (2014a, p. 273, see also Schilling & Cruse, 2012) claims that the body model may provide a common ground for social cognition: “[…] on a certain level of functional granularity, this type of core representation [i.e., the body model] might also describe the generic, universal geometry which is shared by all members of a biological species.” Accordingly, Gallese and Metzinger (2003, p. 550) argue elsewhere that one neural network, viz. the mirror neuron system, underlies both "an internal model of reality" as well as a "shared action ontology".[13] As I take it, this means that the body model can be exploited for both one's own grip on the world, as well as yielding the foundation to share this grip. This 'shared grip' then can be seen as a baseline of similarity which underlies some kinds of social understanding. Indeed, sharing parts of the body model is described as "a new window into the social world." (Metzinger, 2014a, p. 273)

7.3.2. Sharing Representations – The Case of Mirror Neurons

The idea of shared representations as a basis for understanding others is by no means a new one. It is reminiscent of the argument from analogy (see chapter 1.1.2.), in that it assumes that one's own resources are used to understand another person. Further, shared representations are related to the notion of simulation. In fact, shared representations gained prominence when the debate about the mirror neuron system and its potential role in both executing own and understanding others' actions and emotions got off. Both accounts thus assume that the same representations or models that are used for self-related processing are re-used for understanding others. Furthermore, there seems to be solid empirical evidence for the existence of shared representations. Most famously, of course, is the fact that mirror neurons are activated in an agent-neutral manner. Since they fire both when an action is observed and executed, the assumption that they contain shared information is not far to seek.

Kilner, Friston and Frith (2007a, 2007b) elaborate on how a PP perspective can illuminate the role of the mirror neuron system in action understanding. To see how, recall that it has been claimed that the mirror neuron system enables the inference of intentions. Hamilton and Grafton (2007a) agree with this assumption and aim to show that mirror neurons indeed encode more than kinematic parameters of a

[13] To have an ontology, to briefly clarify, here means the same as possessing a model of reality.

particular action.[14] In a compelling study, they showed participants several videos of another person opening and closing a box by sliding a lid. In the course of this, participants either observed the same *outcome* (e.g., 'box open') that was achieved by different kinematic movements ('sliding the lid to the right to open the box' vs. 'sliding the lid to the left to open the box) or they observed the same *movement* (e.g., 'sliding the lid to the right') which resulted in different outcomes ('box open' vs. 'box closed'). In order to explain their results, Hamilton and Grafton state that actions can be depicted in a hierarchy of levels of motor representation, ranging from the muscle level to the kinematic (e.g., sliding), the object-goal level (e.g., sliding the lid) to the level of the outcome (e.g., closing/opening the box) (Hamilton & Grafton, 2007a, 2007b). Since mirror neuron activity is only found when the outcome of an action is the same, it is concluded that mirror neurons only encode the result or goal-representation of an action, and not specific kinematic features:

> [t]herefore, these brain regions contain populations of neurons that encode observed action outcomes, regardless of the kinematic parameters of the action. These results support a hierarchical model of action understanding […] in which a cascade of visuomotor processing in parietal and frontal regions allows us to understand the goals and outcomes of other people's actions. (Hamilton & Grafton, 2007a, pp. 1166–1167)

At this point, Kilner, Friston and Frith pick up the idea of a hierarchy of motor actions and claim that a predictive coding account of mirror neurons gives a plausible explanation for how they encode both goals and intentions of actions[15] by the inversion of generative models along the cortical hierarchy. In order to understand the represented goals and intentions of another individual's actions, the system faces the same inverse problem it is faced with in general. Incoming sensory signals carry information about the kinematic of an action, but not about the goal or intention of this action.

[14] In doing so, the authors aim to prove wrong the criticism of Jacob and Jeannerod (2005) who claim that mirror neurons are merely able to encode motor intentions, but not prior, social, and communicative intentions.

[15] Intentions are described as short-term goals that are necessary to achieve long-term goals (Kilner, Friston & Frith, 2007a, p. 162).

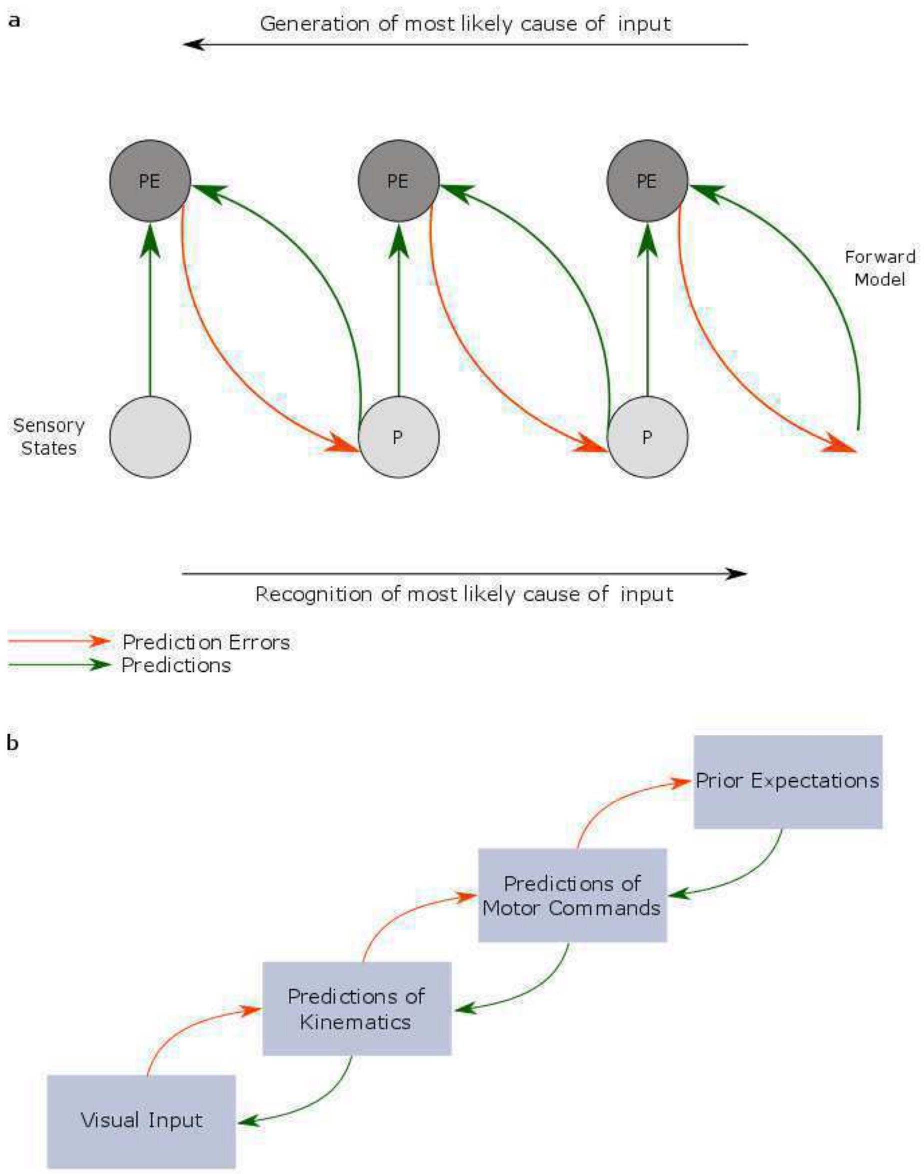

a
Generation of most likely cause of input
PE
PE
PE
Forward Model
Sensory States
P
P
Recognition of most likely cause of input
Prediction Errors
Predictions
b
Prior Expectations
Predictions of Motor Commands
Predictions of Kinematics
Visual Input

Fig. 6 Action observation

a, Given a forward model that is furnished with prior expectations, predictions about the most likely cause of input are generated (light grey circles ‘P’). These are sent down the hierarchy via backward connections (green arrows) and are constantly updated by error signals (grey circles ‘PE’). Prediction errors are conveyed via forward connections (red arrows). By inversion of this process, the most likely cause of visual input can be inferred (Figure adapted from Kilner, Friston & Frith, 2007b). This is described in more detail in Fig. 6 b. ***b***, Given prior expectations about another person’s goal-representation, predictions of underlying motor commands are generated, which in turn enable a prediction of the more fine-grained kinematics underlying an action. These predictions are generated on the basis of the observer’s own motor system and are propagated down the hierarchy via backward connections (green arrows). Finally, there is a comparison between prediction and actual input. The mismatch between prediction and signal is conveyed by forward connections (red arrows) as an error signal and functions to update predictions of kinematics and motor commands. Based upon this update, the represented goal of the other person’s goal can be inferred.

This is what we can call the *social inverse problem*: there is no one-to-one mapping between cause and effect when trying to infer why (cause) another person behaved in a special way (effect).

The motor command, goal or intention that caused the observable action can be manifold, which is why a simple inversion of a forward model is insufficient to infer the most likely cause. This provokes the question the authors pursue in the paper, namely: If mirror neurons enable social inference, how exactly do they do so? Kilner and colleagues claim that a predictive coding account of the mirror neuron system not only offers a solution to what I called the social inverse problem, but furthermore suggests ways to empirically test its neural implementation, making it more powerful and computationally less complicated than earlier attempts to couch action understanding in terms of forward models (e.g., Blakemore, Wolpert & Frith, 2000; Wolpert, Doya & Kawato, 2003, see also chapter 6.3.3.).

The way the mirror neuron system functions during action observation to infer the underlying intention is described as prediction error minimization at all levels of the motor hierarchy (intentions, goals, kinematics, motor commands). On the basis of an expectation of a particular goal-representation that underlies the observed movements, a prediction of motor commands that caused the movement is generated. Given the motor commands, kinematics can be predicted and compared to the actual sensory input. The mismatch between this model and the representation of the visual information is propagated up the hierarchy to update the prediction about the underlying motor commands until the error is minimized to its predicted size. Thus,

> [b]y minimizing the prediction error at all the levels of the MNS, the most likely cause of the action will be inferred at all levels (intention, goal, motor and kinematic). This approach provides a mechanistic account of how responses in the visual and motor systems are organized and explains how the cause of an action can be inferred by its observation. (Kilner, Friston & Frith, 2007a, p. 162)

Crucially, the basis of predictions of possible causes to observed effects in others are generative models about sensory effects of one's own movements, which is in accordance with the general idea of shared representations.

This PP model of the mirror neuron system is a compelling example of how shared representations may contribute to understanding others. There is, in sum, good reason to assume that shared representations play a major role in social cognition.

However, these ideas do not come without doubt, as will become obvious in the next section.

7.3.3. Problems with Shared Representations

In order to understand other people *as* other people, one needs to be able to distinguish them from oneself. A self-other distinction thus is imperative for social understanding. In combination with the claim that shared representations are central to social cognition, this leaves us with a quite delicate predicament: If both shared representations and self-other distinction are hallmarks of social cognition, how can both be achieved at the same time? The so-called *naked problem* (De Vignemont, 2014b; Jeannerod & Pacherie, 2004) describes exactly this problem: Shared content, on the one hand, must not be 'clothed' with any information that is solely self-specific. If it were so, the content would only apply to myself, but not to other people who are different from me in many respects (e.g., body size, motor repertoire, prior experience etc.). The content of shared representations thus must be so broad and unspecific that it potentially could be used for my own and another person's processing. However, on the other hand, if content is indeed naked and applied to both self and other, then these representations cannot code for the agent and thus cannot deliver any information about who did the action, who felt the emotion, or who saw the cat. If there is no agent-specific information contained, these shared representations cannot be the whole story about understanding others. This is exactly what has been criticized about the mirror neuron system and its explanatory power when it comes to higher-level skills like action understanding, imitation learning, or empathy. Newen (2015, p. 4), to give just one out of many examples, argues that mirror neurons cannot form the basis for social understanding:

> Why are mirror neurons not an essential part of understanding others? They represent a type of action or emotion that is independent from a first- or third-person perspective; but the distinction between self and other is an essential part of understanding others.

The basic criticism, thus, is that shared representations implemented by the mirror neuron system yield no self-other distinction. De Vignemont (2014b) picks up this problem and applies it to the case of body ownership and emotions. Indeed, as proof of shared body representations it has been shown that there are overlapping neural systems which are activated both when a person feels, touches or merely observes another person being touched (Singer et al., 2004). In the rare case of mirror touch synesthesia, individuals even feel the touch on their own skin when they see others being touched (Blakemore, 2005). One possible conclusion is the following: "It thus follows from the existence of naked body representations that further processes are needed to discriminate between one's body and other people's bodies." (De Vignemont, 2014b, p. 130) The naked problem therefore involves the problem of self-other distinction; only when representations 'wear' distinctive features and are not 'naked', they can yield a distinction between self and others.

This leads to the second main problem I see with shared representations and which I will call the *genuineness problem*. The issue with any theory that takes an individual's resources as crucial for social cognition is that in order to understand another person *as* another person, to understand her actions as *her* actions and not mine, to grasp *her* emotional states and intentions, it will not be sufficient to simply apply one's own take on a situation. I dub this the 'genuineness problem' because its solution seems to require to use information that is (at least partially) *genuinely* about the other person. The genuineness problem expresses two complications.

First, it leads to the possibility that self and other are confused. While this seems a somewhat absurd problem at first glance (why would I ever mistake you for me?), phenomena such as emotional contagion (e.g., in a mass panic; De Vignemont & Singer, 2006) or neurological conditions such as somatoparaphrenia (in which the patient denies ownership of body parts or her whole body, and often claims that her respective limb belongs to another person; Fotopoulou et al., 2011) are proof that self and other can indeed be mixed up. In the second place, if genuineness is lacking, we run at risk of merely understanding what we would do in a specific situation, but not what the other person is actually doing. We would thus be stuck in a radically egocentric world and end up understanding no one but ourselves.

Another strand of criticism focuses on accounts of mirror neurons which contend that they enable understanding the intention of an observed action. Jacob and Jeannerod (2005) state that the mirror neuron system cannot be said to disambiguate equivocal stimuli and is thus inappropriate to underlie action understanding. This relates to their claim that mirror neurons merely decode motor information and *nothing* above this level (i.e., neither prior nor social intentions, Jacob & Jeannerod, 2005, p. 22). To clarify, they provide the example of Dr Jekyll and Mr Hyde. While the former is a surgeon, the latter is a sadist. Both persona are caught in the same body. If someone were to observe the performance of a motor action (cutting a person with a scalpel) – so they argue – one would not be able to differentiate whether the intention of the action is to *hurt* or *cure* the patient (i.e., whether Dr Jekyll or Mr Hyde is executing the action). This is because the mirror neuron system is assumed to respond in the exact same way when presented with the exact same movement pattern. They conclude that

> [b]y matching them [i.e., the observed actions] onto his own motor repertoire, an observer simulates the agent's movements. Simulating the agent's movements might allow an observer to represent the agent's motor intention. We surmise that it will not allow him to represent the agent's social intention. (ibid., p. 23)

In other words, while mirror neurons may code motor intentions, they do not enable the understanding of personal-level intentions, or what they call 'social intentions'. I will call this the *disambiguation problem*.

In what follows, I will scrutinize each problem and aim to suggest ways to alleviate them.

7.3.4. Contextual Clothing

Finding a solution for the naked and disambiguation problem will crucially depend on one central feature of the PP architecture, namely context-sensitivity. For a more detailed account, we have to make a detour and go back to the level of 1sE, the theory of neural reuse (chapter 7.2.2.) and its combination possibilities with PP. Recall that one central aspect of neural reuse is to view the brain in a network-manner. Of course, describing the basic brain architecture in an interactive network-way is by no means new in the cognitive sciences. However, it will still be a crucial part in finding a solution to our problems.

Additionally to the bare claim that the brain architecture is built in a network-manner, Anderson provides several additions which will turn out to shed more light on central questions of social cognition, specifically on how individuals may distinguish self and other. First, it is emphasized that a variety of modulatory factors influence effective neural connectivity – at different spatial and temporal scales.

These "multi-scale dynamics" (Anderson, 2015, p. 9) range from genetic dispositions to hormonal influences. Secondly, the interplay of bottom-up and top-down effects as well as external (i.e., incoming stimuli) and internal (i.e., prior processing stages) influences should be considered as determining the current neural state and connectivity. This highly complex and dynamic processing is seen as a challenge to any theoretical account of brain function and it is thus proposed that

> […] we need to develop models of explanation that allow for the possibility of top-down and bottom-up mutual constraint, in which both local and global function are synchronically co-determined by the dynamic coupling between elements at various spatial levels of organization. (ibid., p. 10)

It becomes especially clear in this quotation that Anderson's account shows quite a striking resemblance to PP, for both theories depict neural processing as determined by rapidly changing and manifold factors. In order to capture this idea, the notion of TALoNS (transiently assembled local neural subsystems) is introduced. TALoNS are defined as neural subsystems that possess intrinsic functional dispositions, but whose functional role unfolds in the interaction with other neural subsystems. Their functional properties are determined both by their internal structure and effective connectivity (cf. ibid.). Importantly, in accordance with the quotation above, TALoNS are formed and determined by a variety of factors at multiple scales. This flexibility of formation exhibits, according to Clark (2016, p. 151), another strong connection between neural reuse and PP, since

> PP implements just such a fully flexible cognitive architecture and offers a picture of neural dynamics that is highly sensitive, at multiple timescales, both to varying task-demands and to the estimated reliability (or otherwise) of specific bodies of top-down expectation and bottom-up sensory input.

Most obvious is the emphasis on how both intrinsic as well as extrinsic factors shape the response profile of neural assemblies and thus determine their functionality. Recall that in PP, experience will shape priors and predictions, and will influence processing throughout the hierarchy. Therefore, as Clark (2016) argues, PP integrates key insights from the perspective of neural reuse, most importantly the centrality of context-sensitivity in the hierarchical processing model.

Due to the flexible precision-weighting of top-down and bottom-up influences, any prediction may change how incoming input is interpreted and *vice versa*. Drawing on Friston and Price (2001), Clark (cf. 2016, p. 142) argues that the resulting insight is that “the representational capacity and inherent function of any neuron, neuronal population, or cortical area is dynamic and context-sensitive” and that “neuronal responses, in any given cortical area, can represent different things at different times.” In this way, PP inhabits the central feature of neural reuse which is interactive differentiation. For neural assemblies encode functionally differentiated response profiles, and when these interact with each other, “transient task-specific processing regimes (involving transient coalitions of neural resources) […] emerge as contextual effects.” (ibid.) These are able to shape the way information is processed. Thusly formulated, it becomes obvious just how central contextual effects are for processing – and that they are deeply rooted in the very way our brains work. What emerges here is thus not only a promising and fundamental relation between 1sE and 2sE, but also a possible solution to our problems.

To see how context-sensitivity may alleviate the naked problem, we have to consider the possibility that the distinction between self and other – in some instances – can be seen as a context effect. The general idea is that empirical priors influence information processing even before stimuli occur. Thus, it is plausible to assume that contextual cues already trigger priors that are ‘other-related’: “The context within which sensory stimulation is perceived will therefore influence priors and high level priors, resulting in self-other distinctions being dependent on expectations prior to the presentation of a self or other stimulus.” (Apps & Tsakiris, 2014, p. 92)

Does this mean that context does all the heavy lifting for a self-other distinction and that we can get rid of the concept of shared representations from here on? If context already disambiguates who is the agent, why not assume that there are separate self-related and other-specific representations, which are triggered and used depending on what context demands? In other words, why share representations if there are representations tailored to the other person? While these appear to be fair questions, let us consider them under the rationale of efficiency.

It has been claimed that the predictive brain is a lazy brain in the sense that generative models will always search for the most efficient way to solve a task. In understanding another person, it is probably less costly to use information already available to the system, i.e., one’s own predictive models about sensory consequences. The alternative would be to form brand new models for the other person, which appears rather inefficient. However, it should also be clear that we can only use *part* of our own models to understand other agents in order to avoid the naked and genuineness problem. How, then, is the self-other distinction kept alive throughout hierarchical processing? How do I end up understanding the actions or emotions of the other person, instead of acting or feeling myself?

This can be illustrated by the example of action understanding. The contextual information 'other person' should not only recruit those parts of action representations which contain self-related information about the observed action. It should also trigger to set the gain on proprioceptive prediction errors low. This is because only highly precise (descending) prediction errors can elicit actions. When prediction error is weighted low, though, "we are free to deploy the generative model geared to the production of our own actions as a means both of predicting the visual consequences of another's actions and understanding their intentions." (Clark, 2016, p. 158) In other words, predictions about other agents are informed by our own models, but contextual and prior information serves to avoid confusion between self and other. Furthermore, and I will come back to this in more detail shortly, it is important to notice that there is a constant stream of interoceptive information that is and can only be available to the subject. There thus seems to be a rather rigid internal context that already establishes the basis for a self-other distinction (cf. Metzinger, 2004, p. 289). Therefore, if shared representations face the problem of being naked, context-sensitivity provides the appropriate clothing.[16] In this way, component c_3 – self-other distinction is incorporated.

[16] A similar, although much more metaphorical thought that helps to refute this criticism when applied to body representations is given by De Vignemont (2014b). Her point is that because shared body representations always contain information that is too self-specific to be shared, they do not threaten the distinction between self and other. They are, in her words, "[…] Janus-faced. They face inward as representations of one's body and they face outward as representations of other people's bodies." (ibid., p. 135) This conception is important to clarify what it is that can be shared with others, that is, to specify the content of shared representations. De Vignemont argues that only a very coarse and general representation can serve as a basis for sharing, since individual bodies differ considerable in aspects such as size, gender, posture, etc. These specifics are only applicable to bodies of individuals that are very similar to oneself and thus can only be 'facing towards' oneself. However, what De Vignemont (2014a, p.289; 2014b, p. 134) calls the "body map" contains information that is partially to coarse that it applies to both oneself and others, viz., the basic configuration of body parts. This map serves as a functional tool to localize bodily

Let us now tackle the disambiguation problem. Again, the role of context is central, since it can serve to disambiguate the probability of competing models. Contextual information, according to this scheme, not only provides information about the agent of an observed action. It also serves to trigger prior information with which incoming sensory input is met. Hierarchical predictive processing allows the system to disambiguate equivocal stimuli in virtue of activating models that contain information about what the agent already knows. The system is always ready to meet incoming sensory input with a model that contains knowledge about the larger context of the situation. This means that different individuals will trigger different models and that respective prior information will be consulted if available. Those priors will also provide information about the intentions of an individual – given prior knowledge about an individual will inform which of the predictive models about her have the highest probability and likelihood.

The role of context for solving what I have called the disambiguation problem is also put center stage by Kilner and colleagues (2007a, 2007b). They claim that while it may be true that mirror neurons themselves do not function to disambiguate the manifold of potential underlying causes given a unitary sensory information, the fact that the mirror neuron system is *integrated in a larger hierarchy* can[17].

experiences and is said to enable imitation or the experience of vicarious bodily sensations, which both have been claimed to draw upon shared body representations.

[17] Indeed, one counterargument against the claim that mirror neurons do not underlie understanding others draws on the fact that mirror neurons function in a system, hence mirror neuron system (e.g., Cattaneo & Rizzolatti, 2009; Iacoboni & Dapretto, 2006; Rizzolatti & Craighero, 2004). As Jeannerod and Pacherie (2004, pp. 131–132) describe it, this system not only consists of agent-neutral mirror neurons, but also regions containing neurons that do not have bimodal properties and are solely for self-related processing: “The problem of agent-identification, however, is solved by the fact that other premotor neurons (the canonical neurons) and, presumably many other neuron populations as well, fire only when the monkey performs the action and not when it observes it performed by another agent. This is indeed another critical feature of the shared representations concept: they overlap only partially, and the non-overlapping part of a given representation can be the cue for attributing the action to the self or to the other. The same mechanism operates in humans. Neuroimaging experiments where brain activity was compared during different types of simulated actions (e.g. intending actions and preparing for execution, imagining actions, observing actions performed by other people) revealed, first, that there exists a cortical network common to all conditions, to which the inferior parietal lobule (areas 39 and 40), the ventral premotor area (ventral area 6), and part of SMA contribute; and second, that motor representations for each individual condition are clearly specified by the activation of cortical zones which do not overlap between conditions […].” What everyone can agree on is that it is correct that mirror neurons alone do not yield a self-other

Higher levels in the hierarchy process contextual information which contains cues to infer the most likely cause of an observed action given a particular situation. Hence, if visual presentation of a hand holding a scalpel and cutting flesh was presented in a set up that is reminiscent of an operating room, this information makes the intention of curing more plausible than the one of hurting. Contextual cues are processed at higher levels of the hierarchy and are thought to generate predictions about causes that then serve as empirical priors on action observation: "In this scheme, the intention that is inferred from the observation of the action now depends upon the prior information received from a context level." (Kilner, Friston & Frith, 2007a, p. 164) Empirical support for the claims and arguments I have just depicted is, according to the authors, provided by studies that show that the different kinematic patterns of observed movements that lead to the same outcome are not reflected in distinct activations in the mirror neuron system (e.g., Umiltà et al., 2008).

7.3.5. Are Others a Mere Version of Ourselves?

Commenting on the possibility of using self-related information to understand other individuals, Clark (2016, p. 139) states that "[o]ther agents are thus treated as context-nuanced versions of ourselves." However, this strikes me as a too strong statement when considering that shared representations are only one part of the story. For there are numerous brain regions considered crucial for social understanding which do not have mirror properties and are thus unlikely to code information based on shared representations. Barraclough and Perrett (2011), for example, describe how single cell recordings in different brain areas can inform research on social perception. They review a great number of primate studies for both human and non-human animals and find that while there are probably no single neurons which code for the identities of familiar others (e.g., there is no grandmother neuron in area STS), it is likely that there are cell populations which code different identities (cf. ibid., p. 1745).

distinction. This view, however, is rather impoverished, since it considers mirror neurons in isolation.

There are other points which speak against the thought that others are mere versions of ourselves. To see this and to also provide a way to alleviate both the naked and genuineness problem is to scrutinize the multisensory processing which underlies the body model and to ask in which way we can conceive of multisensory integration as yielding a basis for self-other distinction. To see a possible answer, it is important to notice that it is often claimed that the body model serves as a 'reference frame' which functions to compare actual sensory input to predicted input. This idea can be found in Fotopoulou and Tsakiris (2013, p. 1318), who define the body model as describing "diachronic anatomical, postural and visual features of the body, acting as a reference against which current sensory inputs are compared." Together with the claim that bodily properties are represented probabilistically, it becomes clear that the body model can also function as a tool to test whether or not incoming sensory signals are self- or other-related. It can thus be claimed that the body model cannot only be shared, but may also be used as a testing tool for detecting the cause (either me or you) of incoming exteroceptive sensory signals.

In more detail, consider the asymmetry of access claim and its consequences. The fact that an individual has no first-person access to the other's interoceptive and proprioceptive processing and that social cues are mostly visual feeds into the claim that there should be crucial differences in self- and other-processing. Of course, shared representations and mirror properties of several neural systems are crucial 'bridging mechanisms' between individuals. However, they are most probably not the whole story about understanding others, as research suggests. Again, taking into account the role of contextual disambiguation will shed more light on this issue. In general, it can be said that contextual cues trigger specific priors which already provide a self-other distinction, as described before.

More specifically, it is relevant to note that one important context is the perspective or spatial reference frame from which an individual receives sensory input. It has been shown in numerous studies that a stimulus is processed differently depending on whether it is presented from an egocentric or allocentric viewpoint. For example, Chan and colleagues (2004) demonstrate that the extrastriate body area (EBA) exhibits different response profiles depending on whether a bodily stimulus was displayed from an egocentric or allocentric perspective. Similarly, Saxe and colleagues (2005) conducted a study and found evidence for perspective-dependent activation in EBA and primary somatosensory cortex. Accordingly, Fotopoulou and colleagues (2011) argue for a neuropsychological dissociation of a first- and third-person perspective. They ground their claim in a clinical study in which they let somatoparaphrenic patients either look at their affected body part from a first-person perspective or a third-person perspective. In the latter condition, they looked at their affected body part in a mirror and showed significantly higher ownership as compared to the first condition. Importantly, the effect vanished quickly when the perspective was altered and no long-term alleviation of the symptoms of disownership over the body part was reached. This is relevant since it suggests that current sensory input cannot alter the body model in the long run and that it thus draws on 'stored' representations.

There is an interesting potential continuity between levels of embodiment here. Recall that I stated that at the level of 1sE, the body functions as a reference frame and thus can be said to act as a hyperprior for subsequent processing. This point is strengthened when applied to the level of 2sE. Fotopoulou et al. (ibid., p. 3952) claim that such bodily priors "constitute the integration and interpretation of multisensory signals" due to the long-term experience of perceiving one's body from a specific perspective and the simultaneous experience of other somatic stimuli. Metzinger (2004, p. 289) remarks that

> […] the number of interoceptors and internal transducers is much higher than for any other sensory modality. A large number of tactile mechanoreceptors and proprioceptors, the nociceptors underlying pain experience, visceral receptors, and the activity of the vestibular organ, as well as a whole range of specific brainstem nuclei […] constantly signaling the profile of the internal chemical milieu of the organism to the organism, all contribute to the presentational aspect of the conscious perception of one's own body. […] However, it is important to understand that this process is not a simple upward stream, but one that sets a stable functional context by tying the modulation of cortical activity to the ultimate physiological goal of homeostatic stability […]. This context is an exclusively internal context.

In other words, a constant stream of internal stimuli that is exclusively available to one organism provides a functional setting that has here been described as 'bodily priors'. Taken together with the other results on differences in processing resulting from perspectival alterations, it appears that the reference from which a stimulus is perceived already functions as a crucial contextual cue which biases further processing.

If this is true, it has implications which are central for solving the naked and genuineness problem. Only when uni-modal sensory signals are received from a specific, namely egocentric, reference frame will they be integrated at higher levels and thus lead to a sense of self. This is because they need to be in accordance with multisensory predictions about sensory consequences. These multisensory predictions must be met with incoming signals which are most likely evoked by myself; in other words, high-level predictions that have been generated on the basis of multisensory integration are tested against both external and internal signals – and at least the latter will most certainly not be given by another person. In this way, the multisensory body model provides a self-other distinction; it not only functions to test what is most likely to be me, but also what is most likely to be you.

Further, multisensory predictions depend on bodily priors which demand an egocentric viewpoint. If the body and thus the first-person perspective determine our body model which in turn underlies a sense of self and phenomenal selfhood, this again speaks for the fact that only parts of this model can be shared. This is important for a solution to the genuineness problem. For in order to make sense of other people, information drawn from our own models must be updated with the help of sensory signals (i.e., prediction errors) that are given by the other. Put differently, sharing representations only gets us so far in understanding other individuals. What is needed in order to achieve an adaptive and authentic representation of the other person is the update of our predictions about them via exteroceptive information that genuinely stems from the other person. PP provides just the appropriate mechanism to do so; by precision estimation and weighting, prediction errors (i.e., incoming signals from the other person) will update predictions and priors about the other person. Throughout time, this will enrich our representations and models not just of this one specific individual, but other individuals in general. Thus, even though one basis for predicting another individual has been one's own model, incoming sensory input should quickly revise predictions about the other person. If my own predictions do not fit, prediction errors will serve to update those predictions according to information genuinely given by the other person.

Recall that empirical priors are deemed to be found at intermediate levels of the processing hierarchy since they are formed depending on both top-down and bottom-up information. This also explains why it is easier to understand familiar individuals than unfamiliar ones. The system already has information in the form of priors about that person and will thus be quicker to adapt the appropriate models. Further, this thought feeds back to the role of similarity described in chapter 7.2.1. It may be easier to apply own models to individuals – even those we meet for the first time – whose features are more similar to ours.

To answer the question of whether others are mere versions of ourselves, it can be concluded that to a certain degree, we do treat other individuals as such versions. At the same time, however, a PP stance on the multisensory nature of the body model predicts that these versions are constantly updated and thus contain information that genuinely belongs to the other person, making them different from ourselves.

7.3.6. Interactive Inference 2.0

In chapter 7.2.3., I have introduced the notion of interactive inference (InI) at the level of 1sE and distinguished replicative and complementary interactive inference (RInI/CInI). In principle, interactive inference has been depicted as serving the purpose of disambiguating, confirming or ruling out competing predictive models. Extending the scope to the level of 2sE will show that interacting with other individuals still proves to be a highly efficient means to minimize prediction error and that it will also contribute to alleviate some of the problems of shared representations.

To begin with, InI is a compelling way to solve the disambiguation problem. While contextual cues give the system a clue of what is going on, it will still prove helpful in specific contexts to not merely observe the other and perceptually infer their intentions. Entering an actual interaction should provide all interacting individuals not only with genuine information about the other, but also give them more unambiguous cues to which predictive model has the highest posterior probability. To see this, let us consider the case of replicate interactive inference (e.g., synchronization, mimicry, automatic imitation). At the level of 1sE, replicating the bodily state of the other person leads to more similarity in posture, motor behavior, facial expression, etc. In other words, RInI serves to make the bodies of interacting individuals to be in more similar states. If it is true that higher-order representations are grounded in sensorimotor processes, this should also lead to a more similar representational state of the body model in both individuals. Let me elaborate on this with the example of emotion recognition. Several findings are of central importance here. First, the face is probably the most significant body region with the most relevant information when it comes to social understanding (cf. Farmer, McKay & Tsakiris, 2014, p. 290). Secondly, a great number of studies have shown that the sight of emotional expressions leads to activation in brain areas with mirror properties (e.g., Wicker et al., 2003). Further, people tend to mimic facial expressions of their interaction partners (cf. Chartrand & Lakin, 2013, p. 287). I reviewed some of the research suggesting that mimicry not only occurs ubiquitously, but that it also has striking effects on social relationships. In turn, there is growing evidence that the tendency to mimic is considerably influenced by top-down effects and prior information about the other person (ibid.). Putting these findings together, the following picture emerges: Visual signals of the other person's facial expression (plus contextual information) trigger generative models about the underlying emotional state – this is where shared representations enter the picture. These predictive models serve as a basis for generating proprioceptive predictions – that is, motor commands underlying the facial expression – and also interoceptive predictions which refer to the internal bodily state the person must have been in to give rise to the emotion displayed on their face. Proprioceptive prediction error can be quashed by changing the state of facial muscles ourselves, thusly mimicking the other person. As described in chapter 6.4.2., interoceptive prediction error can also be minimized by actively changing one's internal environment. For example, Seth (2013) claims that emotions occur when prediction errors are cancelled out for exteroception, interoception and proprioception, thus disambiguating multimodal models generated in the insular cortex. The same may be true for emotion recognition; multimodal predictive models about the cause of incoming exteroceptive signals are confirmed or ruled out by quashing proprioceptive prediction error (mimicry) and interoceptive prediction error through IPP and thus the most likely cause of the observed emotion can be inferred.

Mimicry, as a means of replicative interactive inference, is thus a crucial and fast way to enhance this process.

The rationale here is that greater bodily similarity will lead to greater social similarity and facilitate social understanding. Of course, whether or not interactive inference will be deemed a fruitful way to figure out the other person depends on prior beliefs and expectation about the other person. As already mentioned, top-down effects are pervasive and determine whether or not mimicry occurs. However, this fits nicely in the more general framework of PP, since the multidirectional interplay between bottom-up and top-down effects are of central importance. To describe emotional responses to other people in terms of RInI helps to incorporate the component of c_9 – emotions and thus to meet one valid demand of the interactive turn to pay more attention to this matter. Schilbach and colleagues (cf. 2013, pp. 396–397) propose that responding to another person emotionally may facilitate social understanding and social cognitive processing. The stance outlined here is in accordance with this claim, in virtue of RInI being a means to enhance prediction error minimization by interacting and emotionally responding to the other person.

7.4. Third-Order Social Embodiment

Individuals who phenomenally represent themselves as social, or whose social encounters are accompanied with phenomenal states can be classified as 3sE systems. In this last chapter, I will try and tackle the phenomenology of social encounters, an often overlooked issue in the research field of social cognition. In order to shed more light on the different experiences that accompany social situations, I will take the following steps.

(1) First, I will argue that there is a discontinuity between levels of embodiment in the sense that what is to be found at the level of 3sE (viz., the experiential quality of social cognitive states) says little about its underlying computational counterparts described at the level of 2sE. This has already been argued for the case of direct perception and will now also be applied to explicit theoretical inference and high-level simulation. Further, the notions of transparency and opacity are applied to different kinds of experiences, which are claimed to come as a spectrum.

(2) In chapter 7.4.2., I will show how transparency and opacity can be operationalized within the framework of predictive processing of sensorimotor contingencies (PPSCM), after discussing its value for phenomenological descriptions.

(3) Chapter 7.4.3. brings together these considerations by pursuing the claim of Palmer and colleagues (2015) that whether or not mental states carry the property of presence depends on their counterfactual richness.

(4) Lastly, I will introduce the level of 3sE+ in chapter 7.4.4. This level of embodiment refers to opaque social states. It is argued that they deserve an additional level of embodiment, since they exhibit a rare and very specific kind of experience that is not only available to a very limited class of systems, but also displays an additional representational process.

7.4.1. The Phenomenology of Social Encounters

I have argued in chapter 2.2.3. that phenomenology has a limited but crucial role for research on social cognition. While I rejected the notion of direct perception as a causal mechanism for social cognition, I contented that the experience of directly perceiving the mental states of other individuals is indeed often part of our phenomenology. A similar criticism has been spelled out for TT and ST; in chapter 1.5.2., I argued that both theories – and the mindreading debate in general – is rather confused and does not yield useful tools for research. At the same time, we do sometimes experience ourselves theorizing about other people's behavior or trying to consciously simulate their potential causes for a specific action or feeling. In this sense, all three notions (i.e., direct perception, theorizing, and simulation) can be characterized as phenomenal experiences of social encounters.

However, it should be noted that there is a discontinuity between levels of description. For neither of those phenomenal concepts necessarily describes the underlying cognitive process. I have made the case for direct perception in chapter 2.2.3., but we can ask the same question for theorizing and simulation, too: Do we find the underlying representational processes to be sub-personal kinds of theory building and simulation?

Consider the case of simulation. It is anything but certain that sub-personal processes which have been described as simulation do indeed generate the phenomenal experience of simulation. To see that, consider the many studies conducted to explore the possibility that mirror neuron activity constitutes a case of implicit simulation (for a review, see, for example Cattaneo & Rizzolatti, 2009). In most of these studies, it appears that the phenomenal experience that arises has the character of direct perception rather than explicit simulation. This remains, however, a speculative claim since one can hardly find any phenomenological evaluation of the experience of participants in these set ups. Still, recall that phenomenologists claim that while most of our social perception is direct, there are cases in which theoretical inference or high-level simulation are needed in order to understand the other person. This is an important observation and speaks for the fact that there is a *spectrum* of various ways in which individuals experience social situations. These experiences range from the sense of immediately and effortlessly understanding the other person to the explicit construction of theories in order to find reasons for another's behavior. This claim will be of central importance for the following discussion.

In order to fully answer the above named question, let me go back to some of the claims I have made thus far. In chapter 1.4.2., I described that both ST and TT at some point moved towards 'going sub-personal' in the sense that they claimed that theoretical inference and simulation are indeed sub-personal processes. However, I also argued that the conceptual landscape of the mindreading debate is rather confused and has failed to yield useful terminological tools. I thus opted for a fresh start and to find a set of less equivocal terms. Accordingly, I avoided to use the terms of simulation and theoretical inference to describe processes at the levels of 1sE and 2sE, and instead introduced concepts such as interactive inference or embodied social inference. At this point, I propose to substitute the terms of high-level simulation and theoretical inference at the level of 3sE with the notion of *construction*.

The term is supposed to capture the experience of consciously constructing potential reasons for another person's behavior. It is important to notice that construction shall be used as a phenomenological concept, which serves to describe a specific experience of social encounters. To see the difference between the experience of direct perception and construction, consider the following example. You enter your friend's apartment and find her crying, obviously upset and devastated. Most likely, you will immediately know that something is wrong and that your friend is sad. Sensing her sadness and devastation appears as an immediate grasp, no conscious thought will be necessary to find out about her state of mind. However, if said friend is crying so hard she cannot even speak, you may experience yourself to construct reasons for her emotional outburst. Has she lost her job? What would make me cry so hard? Maybe someone died? This is the difference between direct perception and construction; the former does not entail the conscious experience of making an effort to understand the person, while the latter appears as explicitly figuring out the causes of another person's behavior. While the introduction of construction and direct perception as solely phenomenological concepts has the advantage of not confusing the sub-personal with the personal level of description, one question remains; viz., how are these concepts that are described at the level of 3sE related to lower levels of embodiment?

In the original 1-3E framework, the relation between level 2E and 3E seemed rather straightforward in that it was claimed that the specific contents of the unconscious body model are elevated to the level of conscious experience and thus computationally ground phenomenal selfhood (cf. Metzinger, 2014a, p. 274). As we shall see, things will grow more complicated in the case of social cognition. In order to begin to find solutions to how the levels of 2sE and 3sE are related, the following ingredients are needed: The first step is to come back to the claim that transparent and opaque states are not an all-or-nothing phenomenon. Much rather, the experiential variety of social encounters comes as a spectrum. Secondly, the application of the concepts of transparency and opacity are deemed to shed more light on the variety of phenomenal experiences in social encounters and provide conceptual tools to spell out differences between experiential phenomena.

Remember that transparency describes the phenomenal experience of directly perceiving something without being aware of the fact that this experience is brought forth by more complex process. This applies to the direct perception of mental states, too – sometimes, I 'just know' what you are feeling. There is one important point about the experience of directness that should not be missed. As Zahavi (cf. 2011, 548-549 yearonly) points out, we may still experience ourselves as directly perceiving the other person's mind even in cases of 'unsuccessful' social understanding. For example, even if you try to deceive me or if I fundamentally misunderstand what you are saying, I may still have *the experience* of directly perceiving the causes for your behavior. Put differently, although I can get what you say absolutely wrong, I would still experience myself as immediately and effortlessly understanding what you are saying: "There is, so to speak, nothing that gets in the way, and it is not as if I am first directed at an intermediary, something different from the state, and then only in a secondary step target it." (ibid., p. 548)

Opaque social states, however, bring with them the experience of actively constructing reasons for another person's behavior. The process of representation becomes the content of phenomenal representation. It will become obvious that during social encounters, the system oscillates between transparency and opacity and that different kinds of experiences thus appear as graded phenomena.

In what follows, I will once again exploit the PP framework to try and enlighten this topic.

7.4.2. PPSMC, Transparency and Opacity

To begin with, it should be clarified how PP relates to conscious perceptual experience. The main claim that is widely held is that perceptual experience arises when prediction error is sufficiently explained away by top-down predictions. With that assumption, PP reverses a model of perception as a mostly bottom-up mechanism. In contrast to this traditional view, proponents of PP assume that perception is mainly determined by top-down predictions that are updated by error signals (cf. Seth, 2014, p. 4).[18] In this manner, Hohwy (2012) emphasizes that the predictive model which has the highest posterior probability gets to populate consciousness. In his approach, he disentangles consciousness and attention and describes them from a PP perspective. Though distinct phenomena, consciousness and attention are somehow related. This relation, Hohwy claims, becomes much clearer when attention is interpreted as optimization of precision expectations and

[18] Evidence for this claim comes, for example, from studies showing that neural activity is attenuated when a stimulus is repeatedly presented, a phenomenon known as 'repetition suppression' (Summerfield et al., 2008). The rationale is that prediction error is decreasing, so predictability increases. Further, Melloni and colleagues (cf. 2011, p. 1393) conducted an EEG study which showed that expected stimuli are up to 100ms faster available for reportable, conscious experience than unexpected ones. The authors conclude that expectation lowers the threshold for visibility of stimuli.

consciousness seen getting its content from the ‘winning’ hypothesis (i.e., the one that best explains away prediction error).

In Hohwy’s (2012, p. 5) own words:

> The core idea is that conscious perception correlates with activity, spanning multiple levels of the cortical hierarchy, which best suppresses precise prediction error: what gets selected for conscious perception is the hypothesis or model that, given the widest context, is currently most closely guided by current (precise) prediction error.

This quotation reveals a potential relation between consciousness and attention. The role of attention is to pick those models that are most precise and best fit the sensory evidence. These selected models then get to determine perceptual content.[19] While Hohwy focusses on how PP functionally relates to attention and consciousness (see chapter 6.3.2.), Seth picks out ‘perceptual presence’ and objecthood in conscious perception and connects them to PPSMC (see chapter 6.5.1.)

Perceptual presence here refers to how ‘real’ or ‘present’ things appear to be. While the cat on your lap probably seems very real and as something you can interact with, the starry sky at night appears more abstract. Seth’s (2014) original claim is that the degree of presence is determined by the counterfactual richness of a generative model. Counterfactual encoding of causes is thus central for the phenomenal property of realness or presence, in contrast to other accounts of perceptual presence (e.g., Noë, 2004). Remember that this kind of conditional coding arises with active inference, that is, with the encoding of possible actions and their sensory consequences. However, in contrast to SMT, Seth (2014, p. 9) does not contend that sensorimotor contingencies relate to actual actions:

> […] action is not constitutively necessary for perceptual presence: that is, it is possible to have a rich repertoire of counterfactual predictions endowing a high level of perceptual presence to some content, without any of these predictions simultaneously driving action.

In making this comment, the author breaks with the phenactive view of perception. He does so in drawing a rather internalist picture of counterfactual processing; it is

[19] For empirical support of this claim, see Hohwy, 2012, pp. 8–12; Feldman & Friston, 2010.

the conditional sensorimotor *knowledge*, and not the sensorimotor loop *itself* that endows hierarchical models with counterfactuals.

At the same time, active sampling of the world is still of paramount importance. As Friston and colleagues (2012) argue, saccadic eye movements can be viewed as actively testing hypotheses, being determined by priors about which kind of movement would minimize uncertainty of perceptual predictions most efficiently. This means, as is made clear by the authors, that priors have to encode in a counterfactual manner; "in other words, what we would infer about the world, if we sample it in a particularly way" (ibid., p. 2). Using an example, the authors show how this relates to conscious perception. Imagine you are sitting in your garden and some fluttering appears in the periphery of your vision. It is assumed that your internal states change so to represent the cause of this sensation as a bird, which minimizes (sub-personal!) surprise about them. This changed hypothesis then serves as a basis to select those prior beliefs that will steer your gaze to the direction of the fluttering, because this movement promises to reduce uncertainty about the hypothesis. The chosen priors then generate proprioceptive predictions about the visual consequences and your oculomotor system, which are fulfilled by action. This action is to move your head towards the fluttering sensation and to orient your saccades in a way so they confirm the hypothesis (cf. ibid., p. 4).

While what Friston and colleagues present us with is a view of active inference as *confirming* predictions, Seth argues – as I have detailed before – for a broader functionality of active inference. Its function to disambiguate sensory input is related to counterfactual processing, too. Active inference, according to this claim, serves as "an efficient method for extracting and encoding higher-order world-revealing invariants (more so than passive deconvolution), and in doing so effectively separates hidden causes in the world from those that depend on actions (or properties) of the perceiver." (Seth, 2015a, p. 3)

There are many important hints to Seth's conception of perceptual presence and its relation to active inference. First, he points out that invariants are world-revealing in the sense that they make the world appear as something that is *outside of* and *independent from* our own mental machinery (see also Hohwy, 2014b). The second point is related and focuses on the difference between objecthood and image-hood. While the former is the phenomenal experience of an object as part of the external world, the latter refers to the impression that something is the product of our internal processing. Although objecthood often goes along with perceptual presence, the two can come apart. Seth clarifies this by using the example of perceiving a blue sky. While the sky seems perfectly real and part of the world, it does not so much appear as an object. Also, the smell of dog poop triggers the experience of something very real, without necessarily evoking the perception of an object (especially when you managed to step into it). Thus, the degree to which presence and objecthood may depart depends not only on the context, but also on sensory modalities (Madary, 2014). However, it is rather unlikely that the experience of presence of the blue sky or the dog poop – to stick (no pun intended) with the examples – draws on counterfactually-rich action models.

As a consequence, Seth (2015a) revises his stance and asserts that a more fine-grained distinction of presence and objecthood is needed. In this refined view, objecthood heavily draws on counterfactually-rich hierarchical models, meaning that the more potential ways of manipulating an object we have (viz., in virtue of sensorimotor knowledge implemented in counterfactually-equipped models), the more real an object appears. At the same time, objecthood seems to depend on the depth of hierarchical processing and invariant regularities that are represented at higher, more abstract ('deeper') levels of the hierarchy (cf. Hohwy, 2014, p. 123; Seth, 2015a, p. 3).

Another interesting way to conceptualize presence (and hence to differentiate it from objecthood) is to relate it to the notion of transparency. In a comment on Seth's proposal, Metzinger (2014b) argues that perceptual presence inhabits two components: "nowness" and "realness". Realness can be spelled out as phenomenal transparency. Remember that transparency has been defined as a phenomenal property of conscious representational states, which makes it impossible for a system to access the fact that it is representing something. The relation to realness now becomes obvious; if a system does not represent that what it perceives is enabled via a representational process, its perceptual contents seem 'real', i.e., as part of the external world. When transparency is lost, however, perceptual content is experienced as mind- rather than world-related (cf. ibid., p. 123; Seth, 2015a, p. 4).

Seth picks up this conceptual clarification and adds that a loss of transparency could result from failing to distinguish hidden causes in the world from action- and agent-dependent causes. This happens when counterfactual predictions are violated, that is, when predictions that are associated with world-revealing causes are not met. Fulfilling those predictions would afford active inference. Consider the case of afterimages (cf. Seth, 2015a, p. 4). Imagine your gaze is directed at a spotlight, and when you turn your head, you see an afterimage of this spotlight. According to Seth, the 'unrealness' of this afterimage occurs precisely because the counterfactual prediction that eye movements are supposed to fulfill are violated.

A possible way to operationalize opacity, or the phenomenal experience that what someone perceives is not part of the outside-world, but rather generated by one's own mind, is to turn to precision weighting. Seth (ibid., p. 5) speculates that

> when perceiver-related causes (like motor commands generating eye movements) are bound up in the generative models explaining sensory signals, there may be greater *precision weighting* of signals related to these causes, reflecting their importance in explaining away sensory prediction errors.

In other words, when attention (which has been described as precision optimization) selects perceiver-related causes, the construction process itself becomes available for conscious awareness.

After having clarified how the concepts of transparency and opacity may be spelled out in PP terms generally, let us now turn to the more specific case of social cognition.

7.4.3. The Spectrum of Mental Presence

The refined description of objecthood and presence may also apply to the perception of 'mental presence'. The notion of mental presence has been coined by Palmer and colleagues (2015) and refers to the experience that the mental states of others are indeed part of the external world and are not made up by our own minds. The question they ask is to what extend we have perceptual experience of the mental states of other people and their answer supports one of my central claims; viz., that there is most probably a spectrum, or as they put it, a scale of presence which depends on counterfactual richness:

> It may be that mental states associated with a poorer set of counterfactuals are experienced as conceptual associations or explicit knowledge, while those with a richer set (e.g., those that occur during conversation) are experienced as subjectively veridical states of the surrounding perceptual world. (Palmer et al., 2015, p. 384)

There is a lot to unpack in this quotation. Most importantly, the authors try to apply the notion of perceptual presence to the experience of social encounters and to relate it to Seth's theory of PPSMC. In this sense, it is said that the counterfactual richness of representations determines the degree to which we perceive the mental states of others as world- or mind-related.

Further, the authors construct a relationship between long- and short-term causes that are predicted throughout the hierarchy, and local and global features of perception. Observed behavioral patterns are deemed to relate to rather local features operating on lower spatiotemporal scales, while inferred mental states are connected with global perceptual features, operating on larger, more abstract scales of the processing hierarchy. What is important here is that there is a bidirectional relationship between stages of processing, in the sense that higher-order representations contain predictions about local perceptual features (cf. ibid., p. 380). These lower levels in turn update predictions at higher levels in virtue of prediction error signals. Therefore, mental state representations rest not only on the basis of the system's predictions, but are also crucially informed by sensory signals given by the other person.

This is an important point for several reasons. On the one hand, the bi-directional interaction prevents the genuineness problem. On the other hand, the assumption that mental state inference rests on genuine other-information is relevant to the phenomenology of social encounters. For phenomenologists have long argued that what we perceive is indeed the other person and not some simulated model of her or a version of ourselves. Of course, this has to be taken with a grain of salt, given the objections raised against the explanatory scope of phenactivist accounts. The perspective just laid out, however, gives us now a reason to believe that there is some truth to this claim, in the sense that prediction errors which carry genuine other-information have an indispensable role in processing and thus determine what ends up at the level of conscious perception.

Another implication of Palmer et al's (cf. 2015, pp. 380–381) view is that the spectrum of short- and long-term expected causes maps onto a spectrum of perspectival (i.e., agent-related) and perspective-invariant (i.e., world-related) levels of representation. These levels, in turn, correspond with levels of awareness and characterize different experiences, as we shall see now, when we take the next step and take transparency and opacity into account.

I claim that in social encounters, transparent phenomenal states arise if the underlying generative model is counterfactually rich, that is, when there is a large set of potential interactions with the other person. To see why, consider that the increase in possible ways to actively test models leads to an increase of possible ways to reveal world-related causes. If precision weighting additionally attributes enough certainty to these models, causes are experienced as being part of the world and not our minds. Further, as Seth claims, an increase in precision weighting on perceiver-related causes ends up in an increase of awareness of the construction process, hence opacity. In other words, when precision is estimated high for a model which explains incoming signals as being caused by the system itself, this *model of the modelling process* is more likely to reach the level of consciousness. In this case, we perceive the mental state representation of the other person as *constructed* by ourselves.

Additionally, as is also noted by Palmer and colleagues, richness increases with the possibilities of directly acting upon causes. Accordingly, the temporal scale of representations throughout the hierarchy may partially determine the counterfactual richness of generative models. For actions occur at short timescales, higher level models that occur on larger timescales may have less counterfactual richness (cf. ibid., p. 384). Thus, whenever it is possible to interact with another person and grasp their hidden mental states via interactive inference, this happens at a relatively short timescale.

However, when counterfactual predictions about sensory causes are violated during an interaction, perceiver-related causes may gain precision and thus appear as more mind-related. For example, when you do not react the way I predict, I may suddenly begin to construct reasons for your behavior, instead of relying on the sensory signals you send. This can, of course, change often and rapidly during an interaction. We may thus oscillate between different experiences during social encounters, ranging from effortlessly and 'directly' perceiving each other's mental states to finding ourselves in need of construction and explicitly thinking about the reasons for the other's actions.

Taken together, the following picture emerges. Transparent states arise in virtue of counterfactually rich hierarchical models that contain world-related predictions of expected causes. They are perspective-invariant and thus also bring forth the phenomenal experience of immediate, direct perception. The corresponding experience in social encounters may be situations in which we 'just know' what the other person is feeling or planning.

Opacity arises and increases with counterfactual poverty and mind-related expectations that are rather perspectival and world-independent. Opaque mental states give rise to the experience of the agent as the cause of mental content. Further, these states seem to come with an additional level of representation; the representation process itself is explicitly represented and available for conscious awareness. Applied to social understanding, we may experience ourselves as constructing reasons for the other person's behavior.

Accordingly, cases of experienced direct perception can be said to result from transparent social cognitive states, while the subjective experience of explicit construction results from opaque states. With this description, it is possible to keep their phenomenal status, while preventing an equation of phenomenal quality and epistemic mechanism. What distinguishes transparent from opaque social cognitive states is, in sum, the degree to which the representation process itself is represented, as well as the spatiotemporal level at which processing takes place. Further, the mind- or world-relatedness of representations determines phenomenal experience.

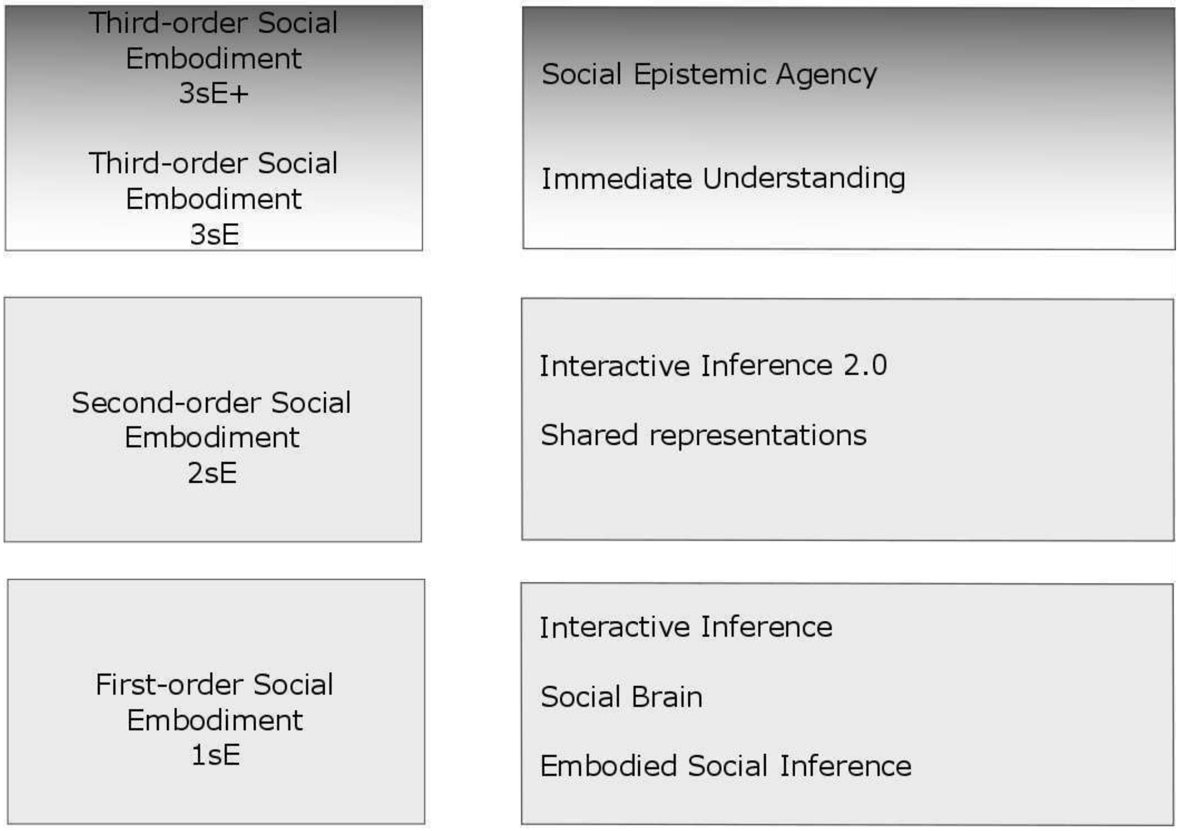
7.4.4. Taking Phenomenal Experience to Another Level – The Case of 3sE+

Additionally to what has been said before, I see a fundamental difference between transparent and opaque phenomenal states in social cognition. Opaque states, it seems to me, exhibit an additional level of representation, for the representation process itself becomes part of the phenomenal experience. In other words, there are some moments when I experience myself as actively constructing the thought process, as explicitly trying to figure out the other person's behavior. This important additional property should be emphasized, since it depicts a fairly special and rare phenomenon, which may also serve to qualify yet another class of social systems. While many animals, for example, probably have transparent social cognitive states in the sense that their interactions with conspecifics are accompanied by some kind of conscious awareness, the explicit representation of reasoning about another individual is probably only available to healthy humans of a specific age. In order to do justice to the specificity of this phenomenon, I introduce the level of "3sE+" which refers to opaque mental states in social encounters.

Transparent and opaque social cognitive states should both certainly be located at the third level of embodiment, since both possess phenomenal properties. This is what has been suggested by Metzinger (cf. 2014a, p. 274) to be the distinctive feature of 2E and 3E systems – to be capable to phenomenally identify oneself with one's body. The resulting properties of phenomenal selfhood or body ownership stem from the experiential quality of immediacy which comes with transparency (cf. ibid., p. 273). This has an important implication, namely that transparency and opacity are not all-or-nothing-phenomena, but that transparency is a property of *any* phenomenal state. This speaks for the depiction of the levels of 3sE and 3sE+ as a spectrum, since the degree to which the representation process is explicitly represented varies, and opacity is a gradually arising property (cf. Metzinger, 2003, p. 358). However, transparent and opaque mental states – at least in the case of social understanding – display two rather different *kinds* of experiences. What is more, their computational counterparts can be differentiated on grounds of counterfactual richness, world- or mind-relatedness and the degree of perspective-invariance. All of this justifies the additional level of 3sE+, also to conceptually draw a clearer distinction.

We can now begin to see how the levels of 2sE and 3sE or 3sE+, respectively, may relate. Whenever interactive inference drives a social encounter and counterfactually rich models are in charge of each interacting person's behavior, the experience of a smooth and effortless understanding arises. However, whenever precision estimation shifts towards counterfactually poorer models that favor agent-related causes, one may find herself constructing reasons for the other person's behavior, maybe theorizing about potential causes or trying to consciously think about what she had done if she was in the other's place.

3sE+ can also be conceptually grasped as exhibiting a kind of 'social epistemic agency'. While most consciously experienced social cognition is automatic and unintentional, 3sE+ experiences can have the additional quality of actively controlling the contents of one's mind. In these cases, one experiences oneself as an epistemic agent. Metzinger (2015) captures this idea for general conscious experiences of mental agency with the terms "M-autonomy" and the "epistemic agent model" (EAM).



Third-order Social Embodiment 3sE+ Third-order Social Embodiment 3sE	Social Epistemic Agency Immediate Understanding
Second-order Social Embodiment 2sE	Interactive Inference 2.0 Shared representations
First-order Social Embodiment 1sE	Interactive Inference Social Brain Embodied Social Inference

Fig. 7 The framework of 1-3sE

The left column describes the levels of social embodiment. In the right column, the respective mechanisms and experiences are listed.

The two related concepts express the specific conscious quality of being in control of one's mental actions (M-autonomy), and controlling one's epistemic relation to the world (EAM). To be an autonomous cognitive and epistemic agent not only involves the ability to select mental contents, but also to be able to inhibit and terminate ongoing conscious thought. Applied to the social realm, episodes of consciously thinking about another person that are actively steered and can be terminated willingly would involve M-autonomy. Experiencing oneself as a subject that thereby stands in a specific epistemic relation to its social world can be described as possessing a *social epistemic agent model* (SEAM).

Taken together, the levels of 3sE and 3sE+ offer a differentiated description of different experiences of social encounters and thusly integrate the component of experiential quality (Fig. 7). The distinction between the two levels of embodiment has been described in terms of transparency and opacity. These concepts have then been operationalized under PP, in the sense that the former are related to perspective-invariant, and the latter to perspective-invariant levels of representation. Additionally, the role of interaction is again put into focus, since representations that contain a richer repertoire of (inter-)action-potentials are thought to bring forth the experience of immediately and effortlessly understanding the other person. The picture of the conscious social mind that emerges here is manifold and captures several phenomenal signatures of being social.

Conclusion

The question I pursued in this work was what kind of philosophical theory is needed to describe the diverse and manifold phenomenon of social cognition in a unifying and comprehensive manner. Throughout Part I of this thesis, I started to tackle this question by listing components of social cognition that need to be considered when theoretically and empirically investigating the phenomenon. Additionally, desiderata for a philosophical theory of social cognition were distilled. In doing so, I rejected several already existing theoretical approaches on the grounds that they do not fulfill the relevant desiderata, or leave out important components, thus drawing an incomplete picture of the phenomenon. In Part II, the aim was to put together a framework that alleviates the problems of former theories and meets the criteria described in Part I. In doing so, I presented two building blocks, 1-3E and PP. These theories have been claimed to not only be compatible, but complementing each other in important ways. The fusion of these two views then formed the basis for my own positive approach, 1-3sE. This framework depicts social cognition as a widely diverse phenomenon and is, because of its hierarchical structure, able to integrate all of its components. It further fulfills the desiderata and can thus be seen as a starting point for a philosophical theory which yields a unifying perspective on social cognition.

Results

The main contribution of this work is the positive account of social cognition, the framework of first-, second-, and third-order social embodiment. It is thought to mainly provide a conceptual basis for further interdisciplinary research and to start building a terminology that can be used by both philosophers and scientists. The conceptual tools it yields and its manifold structure enable researchers to investigate and describe social cognition as an embodied and interactive, but still representational phenomenon. Instead of 'throwing out the baby with the bathwater', it thus finds a middle way between radical approaches and stays on consistent metaphysical territory.

In discussing already existing accounts of social cognition, I have depicted phenactivist and cognitivist views as the radical endpoints of a spectrum of theories on social cognition. The former assume we have direct access to the other person through embodied interaction and go even further in stating that two individuals can form an autonomous cognitive structure. The latter, on the other hand, take as a starting point individual brains that are epistemically marooned from the external and social environment. The consequence is that in order to grasp causes of the behavior of others, they need to be inferred through brain-internal mechanisms. I have worked out several problems that come with both theories. The main common issue is that they are too radical and consequently exclude important components of social cognition, without which a description of the phenomenon is incomplete. In this sense, cognitivism has been found to neglect the influence of embodied interaction and emotions, while phenactivism lacks the power to explain high-level phenomena such as forming explicit theories about the other person. Further, phenactivism has been found to be empirically problematic, since its few empirical studies do not point unequivocally to the fact that interaction constitutes social cognition and the findings can be explained in non-phenactive terms. Additionally, both cognitivist and phenactivist approaches operate on a set of unclear terms that lack descriptive power.

Given that phenactivism and cognitivism come with contradictory metaphysical background assumptions, theories that aim to combine elements from these sides of the spectrum have been diagnosed to be at risk of being inconsistent. At the same time, the idea of putting together insights from both theoretical accounts has been claimed to be highly valuable, but must be approached with caution.

In the search of finding a philosophical framework that circumvents the problems of former theories, but is able to integrate the advantages of them, I found that the theory of first-, second-, and third-order embodiment (1-3E) by Metzinger (2014a) has several properties making it fit for an application to social cognition. First, its hierarchical structure of different levels of embodiment enables a differentiated perspective on social cognition and brings order to the vast diversity of the phenomenon. Secondly, it enables predictions about the relation of levels of description, thus allowing to take into account the phenomenal signature of social cognition while at the same time paying attention to what may ground this phenomenology.

There were three concerns I raised about 1-3E. The first targeted the missing description of the relation between first- and second-order embodiment. It remains unclear, or so I argue, how exactly computational processes are grounded in the physiology and morphology of a system and thus leaves out an important goal of the theory. Secondly, the possible objection was raised that 1-3E tries to combine two contradicting assumptions about the mind in putting together representational with non-representational views. The worry is that we cannot at the same time assume that the mind is fully representational and non-representational. Thirdly, the concepts of transparency and opacity were found to be in need of refinement. This entails the acknowledgement of the fact that probably all phenomenal states possess the property of transparency, since there may always be parts of the representational processes that are not represented, even in opaque states. Further, opacity is seen as a very specific and rare property, which needs a stronger differentiation from transparency, since its possession may qualify a proper class of systems.

Thusly, the framework of 1-3sE aims to overcome these shortcomings for research on social cognition and yield a more differentiated perspective. Concerning the first issue, the concept of embodied social inference (EmSI) has been introduced. EmSI refers to the assumption that social cognitive processes are determined and constrained by the body of an individual. Taken as a concept from the PP scheme, it can further be gathered that predictive models that are used for social cognition are determined by the physiology of an agent. In this sense, a possible relation between the first two levels of embodiment emerges. 1sE as the implementational level depicts the bodily structures that then 'ground' higher level representational processes which will be described at the level of 2sE. 1-3sE also clarifies the second issue of combining representational with non-representational claims. With the help of PP, representationalism can be described as a gradually arising phenomenon, which can already be found at the lowest level of embodiment. So-called action-oriented predictions (Clark, 2016) are interpreted as representations that also entail information about how to engage the body and environment in the cognitive process. Representations, or so I argued, are seen as abstractions from the sensory signal, and thus arise as soon as there is a prediction about this signal in the hierarchy. At lower levels, these representations will be less abstract than at higher levels, since the degree of abstraction increases as one goes up the processing hierarchy. PP here enables a view on social cognition as both embodied and representational, thus integrating phenactive and cognitivist demands, but on a consistent basis.

The third issue is resolved in that an additional level of social embodiment – 3sE+ – is introduced. 3sE+ is meant to capture the specific phenomenal signature of opaque states during which an individual experiences itself as a social epistemic agent. The instantiation of a so-called 'social epistemic agent model' (SEAM) exhibits cases during which one experiences oneself as actively searching for the reasons for another person's behavior. With the introduction of these terms, the difference between transparency and opacity becomes much more apparent, thus doing justice to the rare and specific nature of opaque states.

Beyond this, I have applied the framework of predictive processing to social cognition and found that it provides useful tools for integrating hitherto neglected components of the phenomenon at all levels of embodiment. For the first level of social embodiment (1sE), I discussed the notion of the social brain when viewed from a PP perspective combined with neural reuse. What has emerged here is a picture of the social brain as highly dynamic and interrelated. The traditional concept of functional modules has been replaced by what I called 'social functional fingerprints' (SFF), which describe local neural assemblies that exhibit a specific functional disposition for processing social stimuli. The introduction of these ideas was thought to provide a fresh view on the social brain and enable a new way of approaching empirical work.

I furthermore discussed the notion of interactive inference (InI) and two subtypes of it (replicative interactive inference – RinI, and complementary interactive inference – CInI) as concepts that attribute a strong role to interaction for social cognition. InI allows to re-interpret empirical findings on replicative and complementary behavior during interactions as an embodied and inferential procedure which enhances social cognitive processing. At the level of 2sE, PP enables to alleviate problems that come with the central notion of shared representations. I worked out three major issues, the naked problem, the disambiguation problem, and the genuineness problem. The application of PP yielded smooth ways of tackling these problems, so that the concept of shared representations now appears less problematic.

Furthermore, for 3sE, I discussed how the theory of predictive processing of sensorimotor contingencies (PPSMC) can enlighten the conceptualization of the phenomenology of social encounters. In doing so, it was found that transparent phenomenal states during which one experiences oneself as immediately and effortlessly picking up the other's mental state arise in virtue of counterfactually rich hierarchical models. These contain world-related predictions about causes of the other person's behavior and are rather perspective invariant. Opaque states, on the other hand, have been related to counterfactual poverty and mind-related predictions. This brings forth the experience of actively and consciously constructing reasons for the other person's behavior.

1-3sE has been built in order to give a consistent and comprehensive view on social cognition. However, there is still work to be done, since this framework can only be seen as a starting point. Additional theoretical and empirical investigations need to be conducted in order to fine-tune the features and predictions of the theory.

Future Research

The framework I have proposed invites future research on social cognition to tackle some of the issues I have not been able to detail in this work. The most relevant goals shall be listed here.

- The framework of 1-3sE, as briefly mentioned, functions both as a hierarchical view on a specific phenomenon – which I have exploited in this work – but also as a systems view (Fig. 2). An interesting endeavor would thus be to apply the systems view and work out how social systems can be classified within 1-3sE. It would be possible to find criteria a system has to fulfill in order to be classified as, for example, a 3sE+ system. This would be especially relevant for integrating findings from empirical studies such as the false-belief task. If children are able to pass the test, they can be classified as 3sE+ systems, but before then still count as 3sE systems, given that they have other phenomenal experiences during social encounters.
- Methodologically, it will become more important to implement scenarios that entail real-life interactions. Of course, there are still strong limitations given that studies need to be well-controlled. However, investigating the role of interaction will help to find out whether or not it can actually be said to constitute social cognition.
- This relates to the need of more robust data on both the role of emotional engagement and interaction for social cognition. The interactive turn has already triggered the interest in these phenomena and led a growing number of researchers to look into their influences on social cognitive processing. While this is a much welcomed movement, especially phenactivist should yield more unequivocal results for their claims. A better sense of the role of interaction will also give an answer to the question of whether interactive contexts are profoundly different from non-interactive ones, and at which levels. Is there, for example, a specific neural profile of interactions that may underlie a specific phenomenal signature?
- 1-3sE asserts that differences in sensorimotor processes result in differences of predictive models, which can be shared and exploited for social cognition. From this, we can derive the prediction that large differences of sensorimotor processes between individuals will make social cognitive processes that rely on them more difficult. This has been shown for the case of autism. Cook (2016) argues that since the

kinematics of movements in typical and autistic individuals deviate, they are less likely to resonate with each other. This may be one cause not only for the social impairments that come with autism, but also for the difficulties that typical individuals have in understanding autistic individuals. From the perspective of interactive inference, it can be assumed that processes of replication are disrupted, thus leading to an impaired inference process between individuals. At the level of 2sE, it can be hypothesized that predictive models that are built on the basis of an individual's own motor repertoire are too different to support a stable inference mechanism. Future research should investigate at which level impairments occur and cause impaired social interactions between autistic and neurotypical individuals.

- How do individual differences influence social cognition? This question needs to be broken down into several sub-issues and more attention in future research. As described before, in the case of autism it has been hypothesized that differences in kinematic profiles between individuals with autism and typically developed individuals are one source of problems in social understanding (Cook, 2016). It has further been shown that similar motor experience of individual enhances imitative behavior (Kilner, Hamilton & Blakemore, 2007). These findings can serve as a starting point to examine how important individual similarity and differences are for social cognitive processing. At the neural level, differences in precision weighting could influence the tendency to imitate. This would be predicted by the claim that precision optimization is a leading component in automatic imitation.
- It would further be desirable if empirical studies involved phenomenological investigations of the kind of experience that a specific set-up triggers. This way, it would be possible to find out which phenomenal experience relates to which kind of underlying processing. As laid out in previous chapters, studies that focus on simulation are thought to often bring forth the experience of direct perception. It would be highly interesting to see whether this proves correct and would yield insights in how levels of embodiment are related.

Taken together, research on social cognition is moving fast and gets input from a variety of disciplines. The interdisciplinarity of the research field shapes the diversity of incoming information about our target phenomenon, but does not come without challenges. The role of philosophy, in my opinion, is to integrate this information and to find a consistent, theoretical approach on whose basis existing and future findings can be merged into a comprehensive picture. This work has aimed to take a first step towards this ambitious and exciting road.

Bibliography

Adams, F., & Aizawa, K. (2008). *The bounds of cognition.* Malden, MA: Blackwell Pub.

Adamson, L. B., & Frick, J. E. (2003). The still face: A history of a shared experimental paradigm. *Infancy, 4*(4), 451–473. doi: 10.1207/S15327078IN0404_01

Adolphs, R. (2001). The neurobiology of social cognition. *Current opinion in neurobiology*. (11), 231–239.

Adolphs, R. (2003a). Investigating the cognitive neuroscience of social behavior. *Neuropsychologia*. (41), 119–126.

Adolphs, R. (2003b). Is the human amygdala specialized for processing social information? *Annals New York Academy of Sciences*. (985), 326–340.

Adolphs, R. (2003c). Cognitive neuroscience of human social behaviour. *Nature Reviews Neuroscience, 4*(3), 165–178. doi: 10.1038/nrn1056

Adolphs, R. (2006a). How do we know the minds of others? Domain-specificity, simulation, and enactive social cognition. *Brain Research, 1079*(1), 25–35. doi: 10.1016/j.brainres.2005.12.127

Adolphs, R. (2006b). What is special about social cognition? In J. T. Cacioppo, P. S. Visser, & C. L. Pickett (Eds.), *Social neuroscience. People thinking about thinking people* (pp. 269–285). Cambridge, Mass: MIT Press.

Adolphs, R. (2009). The social brain: Neural basis of social knowledge. *Annual Review of Psychology, 60*(1), 693–716. doi: 10.1146/annurev.psych.60.110707.163514

Adolphs, R. (2010a). Conceptual challenges and directions for social neuroscience. *Neuron, 65*(6), 752–767. doi: 10.1016/j.neuron.2010.03.006

Adolphs, R. (2010b). What does the amygdala contribute to social cognition? *Annals of the New York Academy of Sciences, 1191*(1), 42–61. doi: 10.1111/j.1749-6632.2010.05445.x

Adolphs, R., & Anderson, D. (2013). Social and emotional neuroscience. *Current opinion in neurobiology, 23*(3), 291–293.

Adolphs, R., Gosselin, F., Buchanan, T. W., Tranel, D., Shyns, P., & Damasio, A. R. (2005). A mechanism for impaired fear recognition after amygdala damage. *Nature, 433*, 68–72.

Adolphs, R., Tranel, D., & Damasio, A. R. (1998). The human amygdala in social judgment. *Nature, 393*, 470-474.

Bibliography

Adolphs, R., Tranel, D., Damasio, H., & Damasio, A. R. (1994). Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. *Nature, 372*, 669–672.

Ainley, V., Brass, M., & Tsakiris, M. (2014). Heartfelt imitation: High interoceptive awareness is linked to greater automatic imitation. *Neuropsychologia, 60*, 21–28. doi: 10.1016/j.neuropsychologia.2014.05.010

Aizawa, K. (2010). Consciousness: Don't Give Up on the Brain. *Royal Institute of Philosophy Supplement, 67*, 263–284. doi: 10.1017/S1358246110000032

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: role of the STS region. *Trends in Cognitive Sciences, 4*(7), 267–278. doi: 10.1016/S1364-6613(00)01501-1

Amodio, D. M., & Frith, C. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature Reviews Neuroscience, 7*(4), 268–277. doi: 10.1038/nrn1884

Anderson, M. L. (2010). Neural reuse: a fundamental organizational principle of the brain. *Behavioral and Brain Sciences, 33*(4), 245. doi: 10.1017/S0140525X10000853

Anderson, M. L. (2014). *After phrenology: Neural reuse and the interactive brain.* Cambridge, Mass: MIT Press.

Anderson, M. L. (2015). Précis of After Phrenology: Neural Reuse and the Interactive Brain. *The Behavioral and brain sciences*, 1–22. doi: 10.1017/S0140525X15000631

Anderson, M. L., Kinnison, J., & Pessoa, L. (2013). Describing functional diversity of brain regions and brain networks. *NeuroImage, 73*, 50–58. doi: 10.1016/j.neuroimage.2013.01.071

Anscombe, G. (1975). *Intention* (2nd ed.). Oxford: Basil Blackwell.

Apperly, I. A. (2008). Beyond simulation–theory and theory–theory: Why social cognitive neuroscience should use its own concepts to study "theory of mind". *Cognition, 107*(1), 266–283. doi: 10.1016/j.cognition.2007.07.019

Apps, M. A., & Tsakiris, M. (2014). The free-energy self: a predictive coding account of self-recognition. *Neuroscience and biobehavioral reviews, 41*, 85–97. doi: 10.1016/j.neubiorev.2013.01.029

Arnau, E., Estany, A., González del Solar, Rafael, & Sturm, T. (2014). The extended cognition thesis: Its significance for the philosophy of (cognitive) science. *Philosophical Psychology, 27*(1), 1–18. doi: 10.1080/09515089.2013.836081

Auvray, M., Lenay, C., & Stewart, J. (2009). Perceptual interactions in a minimalist virtual environment. *New Ideas in Psychology, 27*(1), 32–47. doi: 10.1016/j.newideapsych.2007.12.002

Auvray, M., & Rohde, M. (2012). Perceptual crossing: the simplest online paradigm. *Frontiers in Human Neuroscience, 6*. doi: 10.3389/fnhum.2012.00181

Avramides, A. (2001). *Other minds. The problems of philosophy.* London, New York: Routledge.

Baron-Cohen, S. (1997). *Mindblindness.* Cambridge, Mass: MIT Press.

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

Barraclough, N. E., & Perrett, D. I. (2011). From single cells to social perception. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 366*(1571), 1739–1752. doi: 10.1098/rstb.2010.0352

Bastiaansen, J., Thioux, M., & Keysers, C. (2009). Evidence for mirror systems in emotions. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1528), 2391–2404. doi: 10.1098/rstb.2009.0058

Becchio, C., Del Giudice, M., Dal Monte, O., Latini-Corazzini, L., & Pia, L. (2013). In your place: neuropsychological evidence for altercentric remapping in embodied perspective taking. *Social Cognitive and Affective Neuroscience, 8*(2), 165–170. doi: 10.1093/scan/nsr083

Bermúdez, J. L. (2003). The domain of folk psychology. *Royal Institute of Philosophy Supplement, 53*, 25–48. doi: 10.1017/S1358246100008250

Berntson, G. G., & Cacioppo, J. T. (2004). Multilevel analyses and reductionism: Why social psychologists should care about neuroscience and vice versa. In J. T. Cacioppo & G. G. Berntson (Eds.), *Social neuroscience series. Essays in social neuroscience* (pp. 107–121). Cambridge, Mass: MIT Press.

Bird, C. M. (2004). The impact of extensive medial frontal lobe damage on 'Theory of Mind' and cognition. *Brain, 127*(4), 914–928. doi: 10.1093/brain/awh108

Blackburn, S. (1992). Theory, observation and drama. *Mind & Language, 7*(1-2), 187–230. doi: 10.1111/j.1468-0017.1992.tb00204.x

Blakemore, S.-J. (2005). Somatosensory activations during the observation of touch and a case of vision-touch synaesthesia. *Brain, 128*(7), 1571–1583. doi: 10.1093/brain/awh500

Blakemore, S.-J., Wolpert, D. M., & Frith, C. (2000). Why can't you tickle yourself? *NeuroReport, 11*(11). R11-16.

Blanke, O., & Metzinger, T. (2009). Full-body illusions and minimal phenomenal selfhood. *Trends in Cognitive Sciences, 13*(1), 7–13. doi: 10.1016/j.tics.2008.10.003

Borchert, D. (2006). *The encyclopedia of philosophy* (No. 7). Detroit [u.a.]: Thomson Gale, Macmillan Reference.

Brothers, L. (1990). The social brain: A project for integrating primate behavior and neurophysiology in a new domain. *Concepts in Neuroscience, 1*, 27–51.

Brown, E. C., & Brüne, M. (2012). The role of prediction in social neuroscience. *Frontiers in Human Neuroscience, 6*, 1–19. doi: 10.3389/fnhum.2012.00147

Bruin, L. de, & Kästner, L. (2012). Dynamic embodied cognition. *Phenomenology and the Cognitive Sciences, 11*(4), 541–563. doi: 10.1007/s11097-011-9223-1

Buccino, G., Lui, F., Canessa, N., Patteri, I., Lagravinese, G., Benuzzi, F., et al. (2004a). Neural circuits involved in the recognition of actions performed by nonconspecifics: An fMRI Study. *Journal of cognitive neuroscience, 16*(1), 114–126. doi: 10.1162/089892904322755601

Buccino, G., Vogt, S., Ritzl, A., Fink, G. R., Zilles, K., Freund, H. J., & Rizzolatti, G. (2004b). Neural circuits underlying imitation learning of hand actions: an event-related fMRI study. *Neuron, 42*(2), 323–334.

Cacioppo, J. T., & Berntson, G. G. (1992). Social psychological contributions to the decade of the brain: Doctrine of multilevel analysis. *American Psychologist, 47*(8), 1019–1028. doi: 10.1037/0003-066X.47.8.1019

Cacioppo, J. T., & Decety, J. (2011). An introduction into social neuroscience. In J. Decety & J. T. Cacioppo (Eds.), *The Oxford handbook of social neuroscience* (pp. 3–9). New York: Oxford University Press.

Calder, A. J., Keane, J., Manes, F., Antoun, N., & Young, A. W. (2000). Impaired recognition and experience of disgust following brain injury. *Nature Neuroscience, 3*(11), 1077–1078. doi: 10.1038/80586

Calvo-Merino, B., Glaser, D. E., Grèzes, J., Passingham, R. E., & Haggard, P. (2004). Action observation and acquired motor Skills: An fMRI study with expert dancers. *Cerebral Cortex, 15*(8), 1243–1249. doi: 10.1093/cercor/bhi007

Carruthers, G. (2008). Types of body representation and the sense of embodiment. *Consciousness and cognition, 17*(4), 1302–1316. doi: 10.1016/j.concog.2008.02.001

Cattaneo, L., & Rizzolatti, G. (2009). The mirror neuron system. *Archives of neurology, 66*(5), 557–560.

Chan, A. W.-Y., Peelen, M. V., & Downing, P. E. (2004). The effect of viewpoint on body representation in the extrastriate body area. *NeuroReport, 15*, 2407–2410.

Chartrand, T. L., & Lakin, J. L. (2013). The antecedents and consequences of human behavioral mimicry. *Annual review of psychology, 64*, 285–308. doi: 10.1146/annurev-psych-113011-143754

Chemero, A. (2009). *Radical embodied cognitive science.* Cambridge, Mass.: MIT Press.

Cheng, Y., Lin, C.-P., Liu, H.-L., Hsu, Y.-Y., Lim, K.-E., Hung, D., & Decety, J. (2007). Expertise modulates the perception of pain in others. *Current Biology, 17*(19), 1708–1713. doi: 10.1016/j.cub.2007.09.020

Churchland, P. M. (1989). *A neurocomputational perspective: The nature of mind and the structure of science.* Cambridge, MA: MIT Press.

Clark, A. (1990). *Microcognition: Philosophy, cognitive science, and parallel distributed processing. Explorations in cognitive science: Vol. 6.* Cambridge, Mass: MIT Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*, 181–253. doi: 10.1017/S0140525X12000477

Clark, A. (2014). *Mindware: An introduction to the philosophy of cognitive science* (Second Edition). New York, Oxford: Oxford University Press.

Clark, A. (2015a). *Embodied prediction.* In T. Metzinger & J.M. Windt (Eds.), Open Mind (7(T)). Frankfurt am Main: MIND Group. doi: 10.15502/9783958570115

Clark, A. (2015b). *Predicting peace - the end of the representation wars - A reply to Michael Madary.* In T. Metzinger & J.M. Windt (Eds.), Open Mind (7(R)). Frankfurt am Main: MIND Group. doi: 10.5502/9783958570979

Clark, A. (2015c). Radical predictive processing. *The Southern Journal of Philosophy, 53*, 3–27. doi: 10.1111/sjp.12120

Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind.* New York, NY: Oxford University Press.

Conant, R., & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science, 1*(2), 89–97.

Cook, J. (2016). From movement kinematics to social cognition: the case of autism. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 371*(1693). doi: 10.1098/rstb.2015.0372

Cross, E. S., Hamilton, A. F., & Grafton, S. T. (2006). Building a motor simulation de novo: Observation of dance by dancers. *NeuroImage, 31*(3), 1257–1267. doi: 10.1016/j.neuroimage.2006.01.033

Cruse, H. & Schilling, M. (2015). *Mental states as emergent properties.* In T. Metzinger & J.M. Windt (Eds.), Open Mind (9(C)). Frankfurt am Main: MIND Group. doi: 10.15502/9783958570436

Damasio, A. R. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society B: Biological Sciences*. (351), 1413–1420.

Damasio, A. R. (2006). *Descartes' error: Emotion, reason and the human brain.* London: Vintage.

De Jaegher, H. (2009). Social understanding through direct perception? Yes, by interacting. *Consciousness and Cognition, 18*(2), 535–542. doi: 10.1016/j.concog.2008.10.007

De Jaegher, H., & Di Paolo, E. A. (2007). Participatory sense-making. *Phenomenology and the Cognitive Sciences*. (6), 485–507. doi: 10.1007/s11097-007-9076-9

De Jaegher, H., & Di Paolo, E. A. (2013). Enactivism is not interactionism. *Frontiers in Human Neuroscience, 6*, 1–2. doi: 10.3389/fnhum.2012.00345

De Jaegher, H., Di Paolo, E. A., & Gallagher, S. (2010). Can social interaction constitute social cognition? *Trends in Cognitive Sciences, 14*(10), 441–447. doi: 10.1016/j.tics.2010.06.009

De Vignemont, F. (2010). Knowing other people's mental states as if they were one's own. In D. Schmicking & S. Gallagher (Eds.), *Handbook of phenomenology and cognitive science* (pp. 283–299). Dordrecht: Springer Netherlands.

De Vignemont, F. (2014a). Acting for bodily awareness. In R. Shapiro (Ed.), *The Routledge handbook of embodied cognition* (pp. 287–295). Routledge.

De Vignemont, F. (2014b). Shared body representations and the 'Whose' system. *Neuropsychologia, 55*, 128–136. doi: 10.1016/j.neuropsychologia.2013.08.013

De Vignemont, F., & Singer, T. (2006). The empathic brain: how, when and why? *Trends in Cognitive Sciences, 10*(10), 435–441. doi: 10.1016/j.tics.2006.08.008

Decety, J., & Lamm, C. (2006). Human empathy through the lens of social neuroscience. *TheScientificWorldJournal, 6*, 1146–1163. doi: 10.1100/tsw.2006.221

Dennett, D. C. (1971). Intentional systems. *The Journal of Philosophy, 68*(4), 87–106.

Dennett, D. C. (1994). Self-Portait. In S. Guttenplan (Ed.), *A companion to the philosophy of mind* (pp. 236–244). Oxford: Blackwell Press.

Dennett, D. C. (2007). Heterophenomenology reconsidered. *Phenomenology and the Cognitive Sciences, 6*(1-2), 247–270. doi: 10.1007/s11097-006-9044-9

Dennett, D. C. (2011). Shall we tango? No, but thanks for asking. *Journal of Consciousness Studies, 18*(5-6), 23–34.

Dennett, D. C. (2013). *Intuition pumps and other tools for thinking:* Norton & Company, Incorporated, W. W.

Descartes, R. (1986). *Meditationes de Prima Philosophia = Meditationen über die Erste Philosophie.* Stuttgart: Reclam.

Descartes, R., & Cottingham, J. (1641/1996). *Meditations on first philosophy: With selections from the objections and replies; a Latin-English edition.* Cambridge, England: Cambridge University Pr.

deVries, W. (2011). *Wilfried Sellars. The Stanford Encyclopedia of Philosophy,* Edward N. Zalta (Ed.) from http://plato.stanford.edu/entries/sellars/.

Di Paolo, E. A., & De Jaegher, H. (2012). The interactive brain hypothesis. *Frontiers in Human Neuroscience, 6*, 1–16. doi: 10.3389/fnhum.2012.00163

Di Paolo, E. A., & Thompson, E. (2014). The enactive approach. In R. Shapiro (Ed.), *The Routledge handbook of embodied cognition.* Routledge.

Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. *Experimental brain research, 91*(176-180).

Dunbar, R. I. (1992). Neocortex size as a constraint on group size in primates. *Journal of Human Evolution, 20*, 469–493.

Dunbar, R. I. (1998). The social brain hypothesis. *Evolutionary Anthropology*, 178–190.

Engel, A. K., Maye, A., Kurthen, M., & König, P. (2013). Where's the action? The pragmatic turn in cognitive science. *Trends in Cognitive Sciences, 17*(5), 202–209. doi: 10.1016/j.tics.2013.03.006

Fadiga, L., Fogassi, L., Pavesi, G., & Rizzolatti, G. (1995). Motor facilitation during action observation: A magnetic stimulation study. *Journal of Neurophysiology, 73*(6), 2608–2611.

Farmer, H., McKay, R., & Tsakiris, M. (2014). Trust in me: trustworthy others are seen as more physically similar to the self. *Psychological Science, 25*(1), 290–292. doi: 10.1177/0956797613494852

Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience, 4*, 215. doi: 10.3389/fnhum.2010.00215

Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature reviews. Neuroscience, 10*(1), 48–58. doi: 10.1038/nrn2536

Fodor, J. A. (1975). *The language of thought. The language and thought series.* Cambridge, Mass: Harvard University Press.

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology.* Cambridge, Mass: MIT Press.

Fodor, J. A. (1992). A theory of the child's theory of mind. *Cognition, 44*(3), 283–296.

Forbes, C. E., & Grafman, J. (2013). Social neuroscience: the second phase. *Frontiers in Human Neuroscience, 7*, 1–5. doi: 10.3389/fnhum.2013.00020

Fotopoulou, A., Jenkinson, P. M., Tsakiris, M., Haggard, P., Rudd, A., & Kopelman, M. D. (2011). Mirror-view reverses somatoparaphrenia: Dissociation between first- and third-person perspectives on body ownership. *Neuropsychologia, 49*(14), 3946–3955. doi: 10.1016/j.neuropsychologia.2011.10.011

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences, 13*(7), 293–301. doi: 10.1016/j.tics.2009.04.005

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature reviews. Neuroscience, 11*(2), 127–138. doi: 10.1038/nrn2787

Friston, K. (2012). Embodied Inference: or "I think therefore I am, if I am what I think". In J. Kriz (Ed.), *The implications of embodiment: Cognition and communication* (pp. 89–125).

Friston, K., Adams, R. A., Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology, 3*, 151. doi: 10.3389/fpsyg.2012.00151

Friston, K., & Frith, C. (in press). A duet for one. *Consciousness and Cognition*. doi: 10.1016/j.concog.2014.12.003

Friston, K., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics, 104*(1-2), 137–160. doi: 10.1007/s00422-011-0424-z

Friston, K., Daunizeau, J., Kiebel, S. J., & Sporns, O. (2009). Reinforcement learning or active inference? *PloS one, 4*(7), e6421. doi: 10.1371/journal.pone.0006421

Friston, K., & Price, C. J. (2001). Generative models, brain function and neuroimaging. *Scandinavian Journal of Psychology, 42*(3), 167–177. doi: 10.1111/1467-9450.00228

Friston, K., & Stephan, K. E. (2007). Free-energy and the brain. *Synthese, 159*(3), 417–458. doi: 10.1007/s11229-007-9237-y

Frith, C. (1999). Interacting minds -A biological basis. *Science, 286*(5445), 1692–1695. doi: 10.1126/science.286.5445.1692

Frith, C. (2007). The social brain? *Philosophical Transactions of the Royal Society B: Biological Sciences, 362*(1480), 671–678. doi: 10.1098/rstb.2006.2003

Frith, C. (2012). The role of metacognition in human social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1599), 2213–2223. doi: 10.1098/rstb.2012.0123

Frith, C., & Frith, U. (2006). The neural basis of mentalizing. *Neuron, 50*(4), 531–534. doi: 10.1016/j.neuron.2006.05.001

Frith, C., & Frith, U. (2012). Mechanisms of social cognition. *Annual Review of Psychology, 63*(1), 287–313. doi: 10.1146/annurev-psych-120710-100449

Frith, U., Morton, J., & Leslie, A. (1991). The cognitive basis of a biological disorder: autism. *Trends in Neurosciences, 14*(10), 433–438

Froese, T., & Di Paolo, E. A. (2011). The enactive approach: Theoretical sketches from cell to society. *Pragmatics & Cognition, 19*(1), 1–36. doi: 10.1075/pc.19.1.01fro

Froese, T., Iizuka, H., & Ikegami, T. (2014). Embodied social interaction constitutes social cognition in pairs of humans: A minimalist virtual reality experiment. *Scientific Reports, 4*. doi: 10.1038/srep03672

Fuchs, T., & Jaegher, H. (2009). Enactive intersubjectivity: Participatory sense-making and mutual incorporation. *Phenomenology and the Cognitive Sciences, 8*(4), 465–486. doi: 10.1007/s11097-009-9136-4

Fulford, K. W. M., Thornton, T., & Graham, G. (2006). *Oxford textbook of philosophy and psychiatry.* Oxford, New York: Oxford University Press.

Furlanetto, T., Cavallo, A., Manera, V., Tversky, B., & Becchio, C. (2013). Through your eyes: incongruence of gaze and action increases spontaneous perspective taking. *Frontiers in Human Neuroscience, 7*. doi: 10.3389/fnhum.2013.00455

Gallagher, S. (2001). The practice of mind: Theory, simulation or primary interaction? *Journal of Consciousness Studies, 8*(5-7), 83–108.

Gallagher, S. (2005). Phenomenological contributions to a theory of social cognition. *Husserl Studies, 21*(2), 95–110. doi: 10.1007/s10743-005-6402-3.

Gallagher, S. (2007). Simulation trouble. *Social Neuroscience, 2*(3-4), 353–365. doi: 10.1080/17470910601183549

Gallagher, S. (2008). Direct perception in the intersubjective context. *Consciousness and Cognition, 17*(2), 535–543. doi: 10.1016/j.concog.2008.03.003

Gallagher, S., & Hutto, D. D. (2008). Understanding others through primary intersubjectivity and narrative practice. In J. Zlatev, C. Shina, & E. Itkonen (Eds.), *The shared mind: Perspectives on intersubjectivity* (pp. 1–18). Amsterdam: John Benjaminds.

Gallagher, S., Hutto, D. D., Slaby, J., & Cole, J. (2013). The brain as part of an enactive system. *Behavioral and Brain Sciences, 36*(04), 421–422. doi: 10.1017/S0140525X12002105

Gallagher, S., & Zahavi, D. (2008). *The phenomenological mind: An introduction to philosophy of mind and cognitive science.* London, New York: Routledge.

Gallese, V. (2001). The 'Shared Manifold' Hypothesis: From mirror neurons to empathy. *Journal of Conciousness Studies, 8*(5-7), 33–50.

Gallese, V. (2004). *The intentional attunement hypothesis. The mirror neuron system and its role in interpersonal relations.* Retrieved June 17, 2011, from http://www.interdisciplines.org/mirror/papers/1/15#_15.

Gallese, V. (2005). Embodied simulation: From neurons to phenomenal experience. *Phenomenology and the Cognitive Sciences, 4*(1), 23–48. doi: 10.1007/s11097-005-4737-z

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain, 119*, 592–609.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences, 2*(12), 493–501.

Gallese, V., & Lakoff, G. (2005). The brain's concepts: the role of the sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology, 22*(3-4), 455–479. doi: 10.1080/02643290442000310

Gallotti, M., & Frith, C. (2013). Social cognition in the we-mode. *Trends in Cognitive Sciences, 17*(4), 160–165. doi: 10.1016/j.tics.2013.02.002

Gardner, H. (1985). *The mind's new science: A history of the cognitive revolution.* New York: Basic Books.

Garfield, J. L., Peterson, C. C., & Perry, T. (2001). Social cognition, language acquisition and the development of the theory of mind. *Mind and Language, 16*(5), 494–541. doi: 10.1111/1468-0017.00180.

Gazzaniga, M. S. (1985). *The social brain: Discovering the networks of the mind.* New York: Basic Books.

Gazzola, V., Aziz-Zadeh, L., & Keysers, C. (2006). Empathy and the somatotopic auditory mirror system in humans. *Current Biology, 16*(18), 1824–1829. doi: 10.1016/j.cub.2006.07.072

Gibson, J. (1977). The theory of affordances. In R. Shaw (Ed.), *Perceiving, acting, and knowing. Toward an ecological psychology* (pp. 67–82). Hillsdale, NJ: Erlbaum.

Gibson, J. (1979). *The ecological approach to visual perception.* Boston: Houghton Mifflin Company.

Goldman, A. (1989). Interpretation psychologized. *Mind & Language, 4*(3), 161–185.

Goldman, A. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading.* Oxford, New York: Oxford University Press.

Goldman, A., & De Vignemont, F. (2009). Is social cognition embodied? *Trends in Cognitive Sciences, 13*(4), 154–159. doi: 10.1016/j.tics.2009.01.007

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories. Learning, development, and conceptual change.* Cambridge, Mass: MIT Press.

Gopnik, A., & Wellman H. M. (1992). Why the child's theory of mind really is a theory. *Mind & Language, 7*(1-2), 145–171. doi: 10.1111/j.1468-0017.1992.tb00202.x

Gordon, R. M. (1986). Folk psychology as simulation. *Mind & Language, 1*(2), 158–171.

Grafton, S. T., Arbib, M. A., Fadiga, L., & Rizzolatti, G. (1996). Localization of grasp representations in humans by positron emission tomography. *Experimental brain research, 112*, 103–111.

Grafton, S. T., Mazziotta, J. C., Woods, R. P., & Phelps, M. E. (1992). Human functional anatomy of visually guided finger movements. *Brain, 115*(2), 565–587. doi: 10.1093/brain/115.2.565

Hamilton, A. F., & Grafton, S. T. (2007a). Action outcomes are represented in human inferior frontoparietal cortex. *Cerebral Cortex, 18*(5), 1160–1168. doi: 10.1093/cercor/bhm150

Hamilton, A. F., & Grafton, S. T. (2007b). The motor hierarchy: from kinematics to goals and intentions. In P. Haggard, Y. Rossetti, & M. Kawato (Eds.), *Sensorimotor foundations of higher cognition* (pp. 381–407). Oxford: Oxford University Press.

Heberlein, A. S., Ravahi, S. M., Adolphs, R., Tranel, D., & Damasio, A. R. (2000). *Deficits in attributing emotion to moving visual stimuli consequent to amygdala damage.* Retrieved from: http://cognet2.mit.edu/library/conferences/paper?paper_id=47234

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology, 57*(2), 243–259.

Herschbach, M. (2008). Folk psychological and phenomenological accounts of social perception. *Philosophical Explorations, 11*(3), 223–235. doi: 10.1080/13869790802239268

Herschbach, M. (2012). On the role of social interaction in social cognition: A mechanistic alternative to enactivism. *Phenomenology and the Cognitive Sciences, 11*, 467–486. doi: 10.1007/s11097-011-9209-z

Hohwy, J. (2010). The hypothesis testing brain: some philosophical applications. In J. Sutton (Ed.), *9th Conference of the Australasian Society for Cognitive Science* (pp. 135–144).

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology, 3*, 1–14. doi: 10.3389/fpsyg.2012.00096

Hohwy, J. (2013). *The predictive mind.* Oxford: Oxford University Press.

Hohwy, J. (2014a). The Self-evidencing brain. *Noûs*, 1–27. doi: 10.1111/nous.12062

Hohwy, J. (2014b). Elusive phenomenology, counterfactual awareness, and presence without mastery. *Cognitive Neuroscience, 5*(2), 127–128. doi: 10.1080/17588928.2014.906399

Hohwy, J. (2015). *The neural organ explains the mind.* In T. Metzinger & J.M. Windt (Eds.), Open Mind (19(T)). Frankfurt am Main: MIND Group. doi: 10.15502/9783958570016

Hohwy, J., & Palmer, C. (2014). Social cognition as causal infernce: Implications for common knowledge and autism. In J. Michael & M. Gallotti (Eds.), *Social objects. Perspectives on social ontology and social cognition* (pp. 176–189). Dordrecht: Springer Netherlands.

Hohwy, J., & Paton, B. (2010). Explaining away the body: Experiences of supernaturally caused touch and touch on non-hand objects within the rubber hand illusion. *PLoS ONE, 5*(2), e9416. doi: 10.1371/journal.pone.0009416

Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: an epistemological review. *Cognition, 108*(3), 687–701. doi: 10.1016/j.cognition.2008.05.010

Hurley, S. (1998). *Consciousness in action.* Cambridge, Mass: Harvard University Press.

Hurley, S. (2008). Understanding simulation. *Philosophy and Phenomenological Research, LXXVII*(3), 755–775. doi: 10.1111/j.1933-1592.2008.00220.x

Husserl, E. (1973a). Zur Phänomenologie der Intersubjektivität. Texte aus dem Nachlass. Erster Teil: 1905 - 1920 I. Kern (Ed.), *Gesammelte Werke.* Haag: Nijhoff.

Husserl, E. (1973b). Zur Phänomenologie der Intersubjektivität. Texte aus dem Nachlass. Zweiter Teil: 1921-1928 I. Kern (Ed.), *Gesammelte Werke.* Haag: Nijhoff.

Husserl, E. (1973c). Zur Phänomenologie der Intersubjektivität. Texte aus dem Nachlass. Dritter Teil: 1929-1935 I. Kern (Ed.), *Gesammelte Werke.* Haag: Nijhoff.

Husserl, E. (1984). *Logische Untersuchungen.* Den Haag: Martin Nijhoff.

Hutto, D. D. (2008). The Narrative Practice Hypothesis: Clarifications and implications. *Philosophical Explorations, 11*(3), 175–192. doi: 10.1080/13869790802245679

Hyslop, A. (2010). Other minds. *The Stanford Encyclopedia of Philosophy.* http://plato.stanford.edu/archives/fall2010/entries/other-minds/.

Iacoboni, M., & Dapretto, M. (2006). The mirror neuron system and the consequences of its dysfunction. *Nature Reviews Neuroscience, 7*(12), 942–951. doi: 10.1038/nrn2024

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., & Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. *PLoS Biology, 3*(3), 529–535. doi: 10.1371/journal.pbio.0030079

Jabbi, M., Bastiaansen, J., Keysers, C., & Lauwereyns, J. (2008). A common anterior insula representation of disgust observation, experience and imagination shows divergent functional connectivity pathways. *PLoS ONE, 3*(8), e2939. doi: 10.1371/journal.pone.0002939

Jacob, P. (2011). The direct-perception model of empathy: A critique. *Review of Philosophy and Psychology, 2*(3), 519–540. doi: 10.1007/s13164-011-0065-0

Jacob, P., & Jeannerod, M. (2005). The motor theory of social cognition: a critique. *Trends in Cognitive Sciences, 9*(1), 21–25. doi: 10.1016/j.tics.2004.11.003

James, W. (1950). *The principles of psychology*. Vol. II (Authorized ed., Vol. 2). New York: Dover Publications.

Järvinen-Pasley, A., Adolphs, R., Yam, A., Hill, K. J., Grichanik, M., Reilly, J., et al. (2010). Affiliative behavior in williams syndrome: Social perception and real-life social behavior. *Neuropsychologia, 48*(7), 2110–2119. doi: 10.1016/j.neuropsychologia.2010.03.032

Jeannerod, M. (1994). The representing brain: Neural correlates of motor intention and imagery. *Behavioral and Brain Sciences, 17*(02), 187. doi: 10.1017/S0140525X00034026

Jeannerod, M., & Pacherie, E. (2004). Agency, simulation and self-identification. *Mind and Language, 19*(2), 113–146. doi: 10.1111/j.1468-0017.2004.00251.x

Johansson, P. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science, 310*(5745), 116–119. doi: 10.1126/science.1111709

Johnson, S. (2002). Selective activation of a parietofrontal circuit during implicitly imagined prehension. *NeuroImage, 17*(4), 1693–1704. doi: 10.1006/nimg.2002.1265

Kennedy, D. P., & Adolphs, R. (2012). The social brain in psychiatric and neurological disorders. *Trends in Cognitive Sciences*, 559–572. doi: 10.1016/j.tics.2012.09.006

Keysers, C., Kaas, J. H., & Gazzola, V. (2010). Somatosensation in social perception. *Nature Reviews Neuroscience, 11*(6), 417–428. doi: 10.1038/nrn2833

Keysers, C., Wicker, B., Gazzola, V., Anton, J.-L., Fogassi, L., & Gallese, V. (2004). A touching sight. *Neuron, 42*(2), 335–346. doi: 10.1016/S0896-6273(04)00156-4

Kilner, J., Hamilton, A. F., & Blakemore, S. J. (2007). Interference effect of observed human movement on action is due to velocity profile of biological motion. *Social neuroscience, 2*(3-4), 158–166. doi: 10.1080/17470910701428190

Kilner, J., Friston, K., & Frith, C. (2007a). Predictive coding: an account of the mirror neuron system. *Cognitive Processing, 8*(3), 159–166. doi: 10.1007/s10339-007-0170-2

Kilner, J., Friston, K., & Frith, C. (2007b). The mirror-neuron system: a Bayesian perspective. *NeuroReport, 16*(6), 619–623.

Kohler, E., Keysers, C., Umiltà, M. A., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: Action representation in mirror neurons. *Science, 297*(5582), 846–848. doi: 10.1126/science.1070311

Koster-Hale, J., & Saxe, R. (2013). Theory of Mind: A neural prediction problem. *Neuron, 79*(5), 836–848. doi: 10.1016/j.neuron.2013.08.020

Lafrance, M., & Broadbent, M. (1976). Group rapport: Posture sharing as a nonverbal indicator. *Group & Organization Management, 1*(3), 328–333. doi: 10.1177/105960117600100307

Leslie, A. (1987). Pretense and representation: The origins of "Theory of Mind". *Psychological Review, 94*(4), 412–426.

Leslie, A. (1994). Pretending and believing: issues in the theory of ToMM. *Cognition, 50*(1-3), 211–238.

Lewis, D. (1966). An argument for identity theory. *The Journal of Philosophy, 63*(1), 17–25.

Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy, 50*(3), 249–258. doi: 10.1080/00048407212341301

Lieberman, M. D. (2012). A geographical history of social cognitive neuroscience. *NeuroImage, 61*(2), 432–436. doi: 10.1016/j.neuroimage.2011.12.089

Lieberman, M. D. (2013). *Social: Why our brains are wired to connect*. New York: Crown Publishers.

Lipps, T. (1903). *Leitfaden der Psychologie.* Leipzig: Verlag von Wilhelm Engelmann.

Lipps, T. (1907). Das Wissen von fremden Ichen. In T. Lipps (Ed.), *Psychologische Untersuchungen* (pp. 694–722). Leipzig: Verlag von Wilhelm Engelmann.

Ludington-Hoe, S. M., Lewis, T., Morgan, K., Cong, X., Anderson, L., & Reese, S. (2006). Breast and infant temperatures with twins during shared Kangaroo Care. *Journal of obstetric, gynecologic, and neonatal nursing : JOGNN / NAACOG, 35*(2), 223–231. doi: 10.1111/j.1552-6909.2006.00024.x

Lycan, W. G. (1981). Form, function, and feel. *The Journal of Philosophy, 78*(1), 24. doi: 10.2307/2025395

Madary, M. (2014). Perceptual presence without counterfactual richness. *Cognitive Neuroscience, 5*(2), 131–133. doi: 10.1080/17588928.2014.907257

Madary, M. (2015). *Extending the explanandum for predictive processing,* from MIND Group: .

Marr, D. (1982). *Vision: A computational approach.* San Francisco: Freeman & Co.

Mars, R. B., Neubert, F.-X., Noonan, M. P., Sallet, J., Toni, I., & Rushworth, Matthew F. S. (2012). On the relationship between the “default mode network” and the “social brain”. *Frontiers in Human Neuroscience, 6*. doi: 10.3389/fnhum.2012.00189

Matusall, S., Kaufmann, I. M., & Christen, M. (2011). The emergence of social neuroscience as an academic discipline. In J. Decety & J. T. Cacioppo (Eds.), *The Oxford handbook of social neuroscience* (pp. 9–27). New York: Oxford University Press.

Melloni, L., Schwiedrzik, C. M., Müller, N., Rodriguez, E., & Singer, W. (2011). Expectations change the signatures and timing of electrophysiological correlates of perceptual awareness. *The Journal of neuroscience : the official journal of the Society for Neuroscience, 31*(4), 1386–1396. doi: 10.1523/JNEUROSCI.4570-10.2011

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

Meltzoff, A. N. (2007). The 'like me' framework for recognizing and becoming an intentional agent. *Acta psychologica, 124*(1), 26–43. doi: 10.1016/j.actpsy.2006.09.005

Meltzoff, A. N. (2013). Origins of social cognition. In M. R. Banaji & S. A. Gelman (Eds.), *Navigating the social world* (pp. 139–144). Oxford University Press.

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

Meltzoff, A. N., & Moore, K. (1997). Explaining facial imitation: A theoretical model. *Early Development and Parenting, 6*, 179–192.

Menary, R. (2012). Cognitive practices and cognitive character. *Philosophical Explorations, 15*(2), 147–164. doi: 10.1080/13869795.2012.677851

Merleau-Ponty, M., & Smith, C. (1945/2002). *Phenomenology of perception. Routledge classics.* London, New York: Routledge.

Metzinger, T. (2003). Phenomenal transparency and cognitive self-reference. *Phenomenology and the Cognitive Sciences, 2*, 353–393.

Metzinger, T. (2004). *Being no one: The self-model theory of subjectivity.* Cambridge, Mass., London: MIT.

Metzinger, T. (2006). Different conceptions of embodiment. *Psyche, 12*(4), 1-7

Metzinger, T. (2007). Self models. *Scholarpedia, 2*(10), 4174. doi: 10.4249/scholarpedia.4174.

Metzinger, T. (2009). *The ego tunnel: The science of the mind and the myth of the self.* New York: Basic Books.

Metzinger, T. (Ed.) (2010). *Grundkurs Philosophie des Geistes. Intentionalität und mentale Repräsentation.* Paderborn: Mentis.

Metzinger, T. (2014a). First-order embodiment, second-order embodiment, third-order embodiment: From spatiotemporal self-location to minimal selfhood. In R. Shapiro (Ed.), *The Routledge handbook of embodied cognition* (pp. 272–286). Routledge.

Metzinger, T. (2014b). How does the brain encode epistemic reliability? Perceptual presence, phenomenal transparency, and counterfactual richness. *Cognitive Neuroscience, 5*(2), 122–124. doi: 10.1080/17588928.2014.905519

Metzinger, T., & Gallese, V. (2003). The emergence of a shared action ontology: Building blocks for a theory. *Consciousness and Cognition, 12*(4), 549–571. doi: 10.1016/S1053-8100(03)00072-2

Metzinger, T., & Windt, J. (2014). Die phänomenale Signatur des Wissens: Experimentelle Philosophie des Geistes mit oder ohne Intuitionen? In T. Grundmann, J. Horvath, & Kipper J. (Eds.), *Die Experimentelle Philosophie in der Diskussion* (pp. 231–279). Berlin: Suhrkamp.

Metzinger, T., & Windt, J. M. (2015). What does it mean to have an open mind? In T. Metzinger & J.M. Windt (Eds.), Open Mind (19(T)). Frankfurt am Main: MIND Group. doi: 10.15502/9783958571044

Michael, J. (2011). Interactionism and mindreading. *Review of Philosophy and Psychology, 2*(3), 559–578. doi: 10.1007/s13164-011-0066-z

Michael, J., & Overgaard, S. (2012). Interaction and social cognition: A comment on Auvray et al.'s perceptual crossing paradigm. *New Ideas in Psychology, 30*(3), 296–299. doi: 10.1016/j.newideapsych.2012.02.001

Mill, J. S. (1872). *An examination of Sir William Hamilton's philosophy* (4th ed.). London: Longman, Green, Read, and Dyer.

Mitchell, J. P., Mason, M. F., Macrae, C. N., & Banaji, M. R. (2006). Thinking about others: The neural substrates of social cognition. In J. T. Cacioppo, P. S. Visser, & C. L. Pickett (Eds.), *Social neuroscience. People thinking about thinking people* (pp. 63–82). Cambridge, Mass: MIT Press.

Moore, G. E. (1903). The refutation of idealism. *Mind, 12*, 433–453.

Müller, V., & Lindenberger, U. (2011). Cardiac and respiratory patterns synchronize between persons during choir singing. *PloS one, 6*(9), e24893. doi: 10.1371/journal.pone.0024893

Mumford, D. (1992). On the computational architecture of the neocortex: II. The role of cortico-cortical loops. *Biological Cybernetics, 66*, 241–251.

Murray, L., & Trevarthen, C. (1985). Emotional regulations of interactions between two-months-olds and their mothers. In T. M. Field & N. Fox (Eds.), *Social perception in infants* (pp. 177–197). Norwood, NJ: Ablex Publ. Corp.

Newen, A. (2015). *Understanding others.* In T. Metzinger & J.M. Windt (Eds.), Open Mind (26(T)). Frankfurt am Main: MIND Group. doi: 10.15502/9783958570320

Nichols, S., & Stich, S. (2003). *Mindreading: An integrated account of pretence, self-awareness, and understanding other minds.* Oxford, New York: Clarendon; Oxford; Oxford University Press.

Noë, A. (2004). *Action in perception. Representation and mind.* Cambridge, Mass: MIT Press.

Nyqvist, K. H., Anderson, G. C., Bergman, N., Cattaneo, A., Charpak, N., Davanzo, R., et al. (2010). Towards universal Kangaroo Mother Care: Recommendations and report from the First European conference and Seventh International Workshop on Kangaroo Mother Care. *Acta paediatrica (Oslo, Norway : 1992), 99*(6), 820–826. doi: 10.1111/j.1651-2227.2010.01787.x

Ochsner, K. N. (2007). Social Cognitive Neuroscience: Historical development, core principles, and future promises. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social psychology. Handbook of basic principles* (2nd ed., pp. 39–66). New York: Guilford Press.

Ochsner, K. N., & Lieberman, M. D. (2001). The emergence of social cognitive neuroscience. *American Psychologist, 56*(9), 717–734.

O'Regan, K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences, 24*, 939–1031.

Oullier, O., Guzman, G. C. de, Jantzen, K. J., Lagarde, J., & Kelso, J. A. S. (2008). Social coordination dynamics: Measuring human bonding. *Social neuroscience, 3*(2), 178–192. doi: 10.1080/17470910701563392

Overgaard, S., & Michael, J. (2013). The interactive turn in social cognition research: A critique. *Philosophical Psychology*, 1–25 doi: 10.1080/09515089.2013.827109

Palmer, C. J., Seth, A. K., & Hohwy, J. (2015). The felt presence of other minds: Predictive processing, counterfactual predictions, and mentalising in autism. *Consciousness and Cognition*. (36), 376–389. doi: 10.1016/j.concog.2015.04.007

Pezzulo, G., Barsalou, L. W., Cangelosi, A., Fischer, M. H., McRae, K., & Spivey, M. J. (2013). Computational grounded cognition: A new alliance between grounded cognition and computational modeling. *Frontiers in Psychology, 3*. doi: 10.3389/fpsyg.2012.00612

Pickering, M. J., & Clark, A. (2014). Getting ahead: Forward models and their place in cognitive architecture. *Trends in Cognitive Sciences, 18*(9), 451–456. doi: 10.1016/j.tics.2014.05.006

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1*(4), 515–526.

Preston, S. D., & De Waal, Frans B.M. (2002). Emapthy: It's ultimate and proximate bases. *Behavioral and Brain Sciences, 25*(1), 1–71.

Prinz, W. (1990). A common-coding approach to perception and action. In O. Neumann & W. Prinz (Eds.), *Relationships between perception and action* (pp. 167–201). Berlin: Springer Berlin Heidelberg.

Przyrembel, M., Smallwood, J., Pauen, M., & Singer, T. (2012). Illuminating the dark matter of social neuroscience: Considering the problem of social interaction from philosophical, psychological, and neuroscientific perspectives. *Frontiers in Human Neuroscience, 6*, 1–15. doi: 10.3389/fnhum.2012.00190

Pylyshyn, Z. W. (1978). When is attribution of beliefs justified? *Behavioral and Brain Sciences, 1*(4), 592–593.

Quadt, L. (2015). *Multiplicity needs coherence - Towards a unifying framework for social understanding: A Commentary on Albert Newen.* In T. Metzinger & J.M. Windt (Eds.), Open Mind (26(C)). Frankfurt am Main: MIND Group. doi: 10.15502/9783958571112

Richardson, M. J., Marsh, K. L., Isenhower, R. W., Goodman, J. R. L., & Schmidt, R. C. (2007). Rocking together: dynamics of intentional and unintentional interpersonal coordination. *Human movement science, 26*(6), 867–891. doi: 10.1016/j.humov.2007.07.002

Rizzolatti, G., & Arbib, M. A. (1998). Language within our grasp. *Trends in Neurosciences, 21*(5), 188–194. doi: 10.1016/S0166-2236(98)01260-0

Rizzolatti, G., Camarda, R., Fogassi, L., Gentilucci, M., Luppino, G., & Matelli, M. (1988). Functional organization of inferior area 6 in the macaque monkey. *Experimental brain research, 71*, 491–507.

Rizzolatti, G., & Craighero, L. (2004). The mirror neuron system. *Annual Review of Neuroscience, 27*(1), 169–192. doi: 10.1146/annurev.neuro.27.070203.144230

Rizzolatti, G., & Fabbri-Destro, M. (2010). Mirror neurons: From discovery to autism. *Experimental brain research, 200*(3-4), 223–237. doi: 10.1007/s00221-009-2002-3

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996a). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research, 3*(2), 131–141.

Rizzolatti, G., Fadiga, L., Matelli, M., Bettinardi, V., Paulesu, E., Perani, D., & Fazio, F. (1996b). Localization of grasp representations in humans by PET: 1. Observation versus execution. *Experimental brain research, 111*(2), 246–252.

Rizzolatti, G., & Sinigaglia, C. (2010). The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations. *Nature Reviews Neuroscience, 11*(4), 264. Retrieved April 10, 2010, from www.nature.com/reviews/neuro. doi: 10.1038/nrn2805

Ross, D., & Spurrett, D. (2004). What to say to a skeptical metaphysician: A defense manual for cognitive and behavioral scientists. *Behavioral and Brain Sciences, 5*(27), 603–647.

Rowlands, M. (2009). Enactivism and the extended mind. *Topoi, 28*(1), 53–62. doi: 10.1007/s11245-008-9046-z

Saxe, R. (2005). Against simulation: The argument from error. *Trends in Cognitive Sciences, 9*(4), 174–179. doi: 10.1016/j.tics.2005.01.012

Saxe, R. (2006a). Four brain regions for one Theory of Mind? In J. T. Cacioppo, P. S. Visser, & C. L. Pickett (Eds.), *Social neuroscience. People thinking about thinking people* (pp. 41–63). Cambridge, Mass: MIT Press.

Saxe, R. (2006b). Uniquely human social cognition. *Current opinion in neurobiology, 16*(2), 235–239. doi: 10.1016/j.conb.2006.03.001

Saxe, R., Jamal, N., & Powell, L. (2005). My body or yours? The effect of visual perspective on cortical body representations. *Cerebral cortex (New York, N.Y. : 1991), 16*(2), 178–182. doi: 10.1093/cercor/bhi095

Scheler, M. (1912/1973). *Wesen und Formen der Sympathie* (6th ed.). Bern, München: Francke Verlag.

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. doi: 10.1017/S0140525X12000660

Schilling, M., & Cruse, H. (2012). What's next: Recruitment of a grounded predictive body model for planning a robot's actions. *Frontiers in Psychology, 3*. doi: 10.3389/fpsyg.2012.00383

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: Bodies and minds moving together. *Trends in Cognitive Sciences, 10*(2), 70–76. doi: 10.1016/j.tics.2005.12.009

Sellars, W. (1956/1997). *Empiricism and the philosophy of mind.* Cambridge, Mass: Harvard University Press.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences, 17*(11), 565–573. doi: 10.1016/j.tics.2013.09.007

Seth, A. K. (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive Neuroscience, 5*(2), 97–118. doi: 10.1080/17588928.2013.877880

Seth, A. K. (2015a). Presence, objecthood, and the phenomenology of predictive perception. *Cognitive Neuroscience*, 1–7. doi: 10.1080/17588928.2015.1026888

Seth, A. K. (2015b). *The cybernetic bayesian brain - From interoceptive inference to sensorimotor contingencies.* In T. Metzinger & J.M. Windt (Eds.), Open Mind (35(T)). Frankfurt am Main: MIND Group. doi: 10.15502/9783958570108

Seth, A. K., Suzuki, K., & Critchley, H. D. (2011). An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology, 2*. doi: 10.3389/fpsyg.2011.00395

Siewert, C. (2011). Consciousness and intentionality. *The Stanford Encyclopedia of Philosophy.* http://plato.stanford.edu/archives/fall2011/entries/consciousness-intentionality.

Singer, T. (2012). The past, present and future of social neuroscience: A European perspective. *NeuroImage, 61*(2), 437–449. doi: 10.1016/j.neuroimage.2012.01.109

Singer, T., & Lamm, C. (2009). The social neuroscience of empathy. *Annals of the New York Academy of Sciences, 1156*(1), 81–96. doi: 10.1111/j.1749-6632.2009.04418.x

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R., & Frith, C. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science, 303*(5661), 1157–1162. doi: 10.1126/science.1093535

Singer, T., Wolpert, D. M., & Frith, C. (2004). Introduction: the study of social interactions. In C. Frith & D. M. Wolpert (Eds.), *The Neuroscience of social interaction* (pp. xiii–xxvii). Oxford: Oxford University Press.

Slors, M. (2012). The Model-Model of the Theory-Theory. *Inquiry, 55*(5), 521–542. doi: 10.1080/0020174X.2012.716205

Stanley, D. A., & Adolphs, R. (2013). Toward a neural basis for social behavior. *Neuron, 80*(3), 816–826. doi: 10.1016/j.neuron.2013.10.038

Stein, E. (1917/2008). *Zum Problem der Einfühlung.* (Sondermann, M. Antonia, Ed.). Freiburg: Herder.

Stich, S. (1986). *From folk psychology to cognitive science: The case against belief* (2. print). *A Bradford book.* Cambridge/Mass. u.a: MIT Press.

Stich, S., & Nichols, S. (1992). Folk psychology: Simulation or tacit theory? *Mind & Language, 7*(1), 35–71.

Stueber, K. (2013). Empathy. *The Stanford Encyclopedia of Philosophy.* http://plato.stanford.edu/archives/sum2013/entries/empathy/.

Summerfield, C., Trittschuh, E. H., Monti, J. M., Mesulam, M. M., & Egner, T. (2008). Neural repetition suppression reflects fulfilled perceptual expectations. *Nature Neuroscience, 11*(9), 1004–1006. doi: 10.1038/nn.2163

Suzuki, K., Garfinkel, S. N., Critchley, H. D., & Seth, A. K. (2013). Multisensory integration across exteroceptive and interoceptive domains modulates self-experience in the rubber-hand illusion. *Neuropsychologia, 51*(13), 2909–2917. doi: 10.1016/j.neuropsychologia.2013.08.014

Thompson, E. (2010). *Mind in life: Biology, phenomenology, and the sciences of mind.* Cambridge, Mass, London: Belknap.

Thompson, E., & Cosmelli, D. (2013). Brain in a vat or body in a world? Brainbound versus enactive views of experience. *Philosophical Topics, 39*(1). 163-180

Trevarthen, C. (1979). Communication and cooperation in early infancy: A description of primary intersubjectivity. In M. Bullowa (Ed.), *Before speech. The beginning of interpersonal communication* (pp. 321–347). Cambridge: Cambridge Univ. Pr.

Tsakiris, M., Jimenez, A. T., & Costantini, M. (2011). Just a heartbeat away from one's body: Interoceptive sensitivity predicts malleability of body-representations. *Proceedings of the Royal Society B: Biological Sciences, 278*(1717), 2470–2476. doi: 10.1098/rspb.2010.2547

Tsakiris, M. (2008). Looking for myself: current multisensory input alters self-face recognition. *PLoS ONE, 3*(12), e4040. doi: 10.1371/journal.pone.0004040

Tsakiris, M., & Fotopoulou, A. (2008). Is my body the sum of online and offline body-representations? *Consciousness and Cognition, 17*(4), 1317. doi: 10.1016/j.concog.2008.06.012

Umiltà, M. A., Escola, L., Intskirveli, I., Grammont, F., Rochat, M., Caruana, F., et al. (2008). When pliers become fingers in the monkey motor system. *Proceedings of the National Academy of Sciences, 105*(6), 2209–2213. doi: 10.1073/pnas.0705985105

Umiltà, M. A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., & Rizzolatti, G. (2001). I know what you are doing: A neurophysiological study. *Neuron, 31*(1), 155–165.

Varela, F. J., Rosch, E., & Thompson, E. (1993). *The embodied mind: Cognitive science and human experience* (1st ed.). Cambridge, Mass. [u.a.]: MIT Press.

Ward, J. (2012). *The student's guide to social neuroscience.* New York: Psychology Press.

Whiten, A., & Byrne, R. W. (1997). *Machiavellian intelligence II: Extensions and evaluations.* Cambridge, New York, NY, USA: Cambridge University Press.

Wicker, B., Keysers, C., Plailly, J., Royet, J.-P., Gallese, V., & Rizzolatti, G. (2003). Both of us disgusted in my insula: The common neural basis of seeing and feeling disgust. *Neuron, 40*(3), 655–664. doi: 10.1016/S0896-6273(03)00679-2

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. doi: 10.1016/0010-0277(83)90004-5

Wolpert, D. M., Doya, K., & Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences, 358*(1431), 593–602. doi: 10.1098/rstb.2002.1238

Zahavi, D. (2001). Beyond empathy: Phenomenological approaches to intersubjectivity. *Journal of Consciousness Studies, 8*(5-7), 151–167.

Zahavi, D. (2010). Empathy, embodiment and interpersonal understanding: From Lipps to Schutz. *Inquiry, 53*(3), 285–306. doi: 10.1080/00201741003784663

Zahavi, D. (2011). Empathy and direct social perception: A phenomenological proposal. *Review of Philosophy and Psychology, 2*(3), 541–558. doi: 10.1007/s13164-011-0070-3

Zahavi, D. (2012). Empathy and mirroring: Husserl and Gallese. In R. Breeur & U. Melle (Eds.), *Phaenomenologica* (pp. 217–254). Dordrecht: Springer Netherlands.

Zimmer, D. E. (1989). Wilde Kinder. In D. E. Zimmer (Ed.), *Experimente des Lebens* (pp. 21–47). Zürich: Haffmanns Verlag.

Deutsche Zusammenfassung

Diese Arbeitet konzentriert sich auf das Phänomen ‚soziale Kognition' und dessen theoretische Implikationen. Das Forschungsfeld, welches sich mit sozialer Kognition beschäftigt, ist geprägt von einer Vielzahl philosophischer, psychologischer und neurowissenschaftlicher Theorien. Derzeit scheint es zwei verschiedene Intuitionen zu geben, wenn man sich dieses Forschungsfeld genauer ansieht. Die eine Seite behauptet, dass zwischenmenschliches Verstehen ohne großen Aufwand geschieht, dass wir ‚direkt' wahrnehmen, was andere Menschen vorhaben, wie sie sich fühlen, und welche Intentionen ihren Handlungen unterliegen. Auf der anderen Seite steht die Überzeugung, dass die mentalen Zustände des anderen nur über einen Inferenzprozess zugänglich sind, da sie sich nicht eindeutig im Verhalten widerspiegeln. Hinter diesen Intuitionen stehen unterschiedliche Annahmen über die metaphysische Natur des Geistes. Der sogenannte Enaktivismus – welcher die erste Intuition der direkten Wahrnehmung teilt – nimmt an, dass sich der Geist in der Interaktion zwischen Umwelt und Agent manifestiert und daher weder im Innen noch im Außen verortet werden kann. Daraus ergibt sich die Überzeugung, dass soziale Kognition konstituiert von zwischenmenschlichen Interaktionen werde und eine Beschreibung des Phänomens daher nicht auf individuelle Prozesse reduzierbar sei. Im Gegensatz dazu behauptet der Kognitivismus, dass der Geist ausschließlich im Gehirn zu verorten sei und dass jegliche externe Prozesse eine geringe Rolle spielen. Ähnlich stellt sich die Ansicht sozialer Kognition dar; um den anderen zu verstehen, bedarf es lediglich interner Prozesse, während Interaktionen irrelevant seien. Aus diesem derzeitigen Zwiespalt ergibt sich die zentrale Fragestellung dieser Arbeit:

> Welche Art von Theorie wird benötigt, um das Phänomen der sozialen Kognition zu erfassen?

Auf der Suche nach einer Antwort werde ich im ersten Teil der Arbeit drei Forschungsfelder und deren theoretische Annahmen untersuchen. Zuerst betrachte ich eine Debatte, welche in der Philosophie des Geistes, aber auch der Psychologie und später der Neurowissenschaft zu verorten ist. Die sogenannte ‚Mindreading'-Debatte wird hauptsächlich vertreten von Kognitivisten und fokussiert daher Inferenzprozesse. Ich argumentiere, dass sich in dieser Debatte kein zufriedenstellendes Bild sozialer Kognition abzeichnet, da sie wichtige

Komponenten des Phänomens, wie etwa die Rolle von Interaktionen, vernachlässigt. Als nächstes beschreibe und beurteile ich die Diskussion in der Phänomenologie und des Enaktivismus. Hier wird sich herausstellen, dass die vorgeschlagenen Theorien sowohl terminologisch wie auch empirisch nicht gefestigt genug sind, um als Kandidaten zu zählen. Das dritte Feld der sozialen Neurowissenschaft bietet eine Reihe interessanter Forschungsergebnisse, welche dargestellt und hinsichtlich ihrer theoretischen Aussagekraft untersucht werden. Dabei werden auch konzeptuelle und theoretische Probleme, die sich stellen, beschrieben und in Betracht gezogen.

Nach dieser Evaluation existierender Theorien fasse ich zusammen, dass keines der Forschungsfelder eine befriedigende Grundlage für eine philosophische Theorie sozialer Kognition bietet. Aufgrund dieser Kritik schlage ich vor, die fruchtbaren Elemente jeder Theorie zu kombinieren und in ein konsistentes Bild zusammenzufügen. Um dies anzugehen, werden im zweiten Teil der Arbeit zunächst zwei Theorien vorgestellt, welche als Basis für meinen eigenen Vorschlag dienen.

Metzingers (2014) Theorie des „first-, second-, and third-order embodiment" (1-3E) wird in ihren Grundzügen dargestellt und hinsichtlich dreier Kritikpunkte modifiziert. 1-3E dient als Grundgerüst für meine eigene Theorie. Sie erlaubt es, phänomenale Eigenschaften mit komputationalen und physikalischen Ebenen zu verknüpfen und bietet daher einen Rahmen für die Einordnung verschiedener Prozesse eines Phänomens. Als nächstes wird die Theorie „Predictive Processing" (Hohwy, 2013; Clark, 2016) vorgestellt und evaluiert. Diese nimmt an, dass die Hauptaufgabe des Gehirns darin liege, Vorhersagen über seine eigenen Zustände zu machen und diese anschließend mit sensorischen Informationen zu vergleichen. Aus diesem Vergleich werde ein Vorhersagefehler generiert, welcher dann dazu diene, die Vorhersagen zu verbessern, bis ein kohärentes Bild über die Umwelt eines Agenten entstehe.

In einem dritten Schritt wende ich die bisher vorgestellten theoretischen Modelle an und beschreibe meine Theorie des „first-, second-, and third-order social embodiment" (1-3sE). Diese Theorie baut ein hierarchisch organisiertes Grundgerüst, welches verschiedene Beschreibungsebenen besitzt und somit ermöglicht, die Vielzahl an Komponenten sozialer Kognition in einem einzigen theoretischen Rahmen zu vereinen.

Erklärungen gemäß § 6 der Promotionsordnung

Hiermit erkläre ich, **Lisa Anna Quadt**, dass ich im Fach **Philosophie** keine Prüfung an einer Universität oder einer gleichgestellten Hochschule in Deutschland endgültig nicht bestanden habe und mich nicht an einer Universität oder an einer gleichgestellten Hochschule in Deutschland in einem Prüfungsverfahren befinde.

Hiermit erkläre ich, dass die Dissertation selbständig, ohne fremde Hilfe und mit keinen anderen als den darin angegebenen Hilfsmitteln angefertigt wurde, dass die wörtlichen oder dem Inhalt nach aus fremden Arbeiten entnommenen Stellen, Zeichnungen, Skizzen, bildlichen Darstellungen und dergleichen als solche genau kenntlich gemacht sind.

Hiermit erkläre ich, dass die Arbeit noch nicht in gleicher oder anderer Form an irgendeiner Stelle als Prüfungsleistung vorgelegt worden ist.

Datum: 09. Juni 2016

Unterschrift

Lisa Quadt

Zusammenfassung “Levels of Social Embodiment – Towards a Unifying Perspective of Social Cognition”

Diese Arbeitet konzentriert sich auf das Phänomen ‚soziale Kognition’ und dessen theoretische Implikationen. Derzeit scheint es zwei verschiedene Intuitionen zu geben, wenn man sich das Forschungsfeld um soziale Kognition genauer ansieht. Die eine Seite behauptet, dass zwischenmenschliches Verstehen ohne großen Aufwand geschieht, dass wir ‚direkt’ wahrnehmen, was andere Menschen vorhaben, wie sie sich fühlen, und welche Intentionen ihren Handlungen unterliegen. Auf der anderen Seite steht die Überzeugung, dass die mentalen Zustände des anderen nur über einen Inferenzprozess zugänglich sind, da sie sich nicht eindeutig im Verhalten widerspiegeln. Hinter diesen Intuitionen stehen unterschiedliche Annahmen über die metaphysische Natur des Geistes. Der sogenannte Enaktivismus – welcher die erste Intuition der direkten Wahrnehmung teilt – nimmt an, dass sich der Geist in der Interaktion zwischen Umwelt und Agent manifestiert und daher weder im Innen noch im Außen verortet werden kann. Daraus ergibt sich die Überzeugung, dass soziale Kognition konstituiert von zwischenmenschlichen Interaktionen werde und eine Beschreibung des Phänomens daher nicht auf individuelle Prozesse reduzierbar sei. Im Gegensatz dazu behauptet der Kognitivismus, dass der Geist ausschließlich im Gehirn zu verorten sei und dass jegliche externe Prozesse eine geringe Rolle spielen. Ähnlich stellt sich die Ansicht sozialer Kognition dar; um den anderen zu verstehen, bedarf es lediglich interner Prozesse, während Interaktionen irrelevant seien. Aus diesem derzeitigen Zwiespalt ergibt sich die zentrale Fragestellung dieser Arbeit: Welche Art von Theorie wird benötigt, um das Phänomen der sozialen Kognition zu erfassen? Auf der Suche nach einer Antwort werde ich im ersten Teil der Arbeit drei Forschungsfelder und deren theoretische Annahmen untersuchen. Zuerst betrachte ich eine Debatte, welche in der Philosophie des Geistes, aber auch der Psychologie und später der Neurowissenschaft zu verorten ist. Die sogenannte ‚Mindreading’-Debatte wird hauptsächlich vertreten von Kognitivisten und fokussiert daher Inferenzprozesse. Ich argumentiere, dass sich in dieser Debatte kein zufriedenstellendes Bild sozialer Kognition abzeichnet, da sie wichtige Komponenten des Phänomens, wie etwa die Rolle von Interaktionen, vernachlässigt. Als nächstes beschreibe und beurteile ich die Diskussion in der Phänomenologie und des Enaktivismus. Hier wird sich herausstellen, dass die vorgeschlagenen Theorien sowohl terminologisch wie auch empirisch nicht gefestigt genug sind, um als Kandidaten zu zählen. Das dritte Feld der sozialen Neurowissenschaft bietet eine Reihe interessanter Forschungsergebnisse, welche dargestellt und hinsichtlich ihrer theoretischen Aussagekraft untersucht werden. Dabei werden auch konzeptuelle und theoretische Probleme, die sich stellen, beschrieben und in Betracht gezogen.

Nach dieser Evaluation existierender Theorien fasse ich zusammen, dass keines der Forschungsfelder eine befriedigende Grundlage für eine philosophische Theorie sozialer Kognition bietet. Aufgrund dieser Kritik schlage ich vor, die fruchtbaren Elemente jeder Theorie zu kombinieren und in ein konsistentes Bild zusammenzufügen. Um dies anzugehen, werden im zweiten Teil der Arbeit zunächst zwei Theorien vorgestellt, welche als Basis für meinen eigenen Vorschlag dienen. Metzingers (2014) Theorie des „first-, second-, and third-order embodiment” (1-3E) wird in ihren Grundzügen dargestellt und hinsichtlich dreier Kritikpunkte modifiziert. 1-3E dient als Grundgerüst für meine eigene Theorie. Sie erlaubt es, phänomenale Eigenschaften mit komputationalen und physikalischen Ebenen zu verknüpfen und bietet daher einen Rahmen für die Einordnung verschiedener Prozesse eines Phänomens. Als nächstes wird die Theorie „Predictive Processing” (Hohwy, 2013; Clark, 2016) vorgestellt und evaluiert. Diese nimmt an, dass die Hauptaufgabe des Gehirns darin liege, Vorhersagen über seine eigenen Zustände zu machen und diese anschließend mit sensorischen Informationen zu vergleichen. Aus diesem Vergleich werde ein Vorhersagefehler generiert, welcher dann dazu diene, die Vorhersagen zu verbessern, bis ein kohärentes Bild über die Umwelt eines Agenten entstehe. In einem dritten Schritt wende ich die bisher vorgestellten theoretischen Modelle an und beschreibe meine Theorie des „first-, second-, and third-order social embodiment” (1-3sE). Diese Theorie baut ein hierarchisch organisiertes Grundgerüst, welches verschiedene Beschreibungsebenen besitzt und somit ermöglicht, die Vielzahl an Komponenten sozialer Kognition in einem einzigen theoretischen Rahmen zu vereinen.