Mind in practice : a pragmatic and interdisciplinary account of intersubjectivity

Bruin, L.C. de

Mind in Practice

a pragmatic and interdisciplinary account of intersubjectivity

Leon Corné de Bruin

Mind in Practice

a pragmatic and interdisciplinary account of intersubjectivity

Proefschrift ter verkrijging van
de graad van Doctor aan de Universiteit van Leiden,
op gezag van de Rector Magnificus, Prof. mr. dr. P.F. van der Heijden
volgens besluit van het College voor Promoties
te verdedigen op woensdag 29 september 2010
klokke 11.15 uur.

door
Leon Corné de Bruin
geboren te Nijkerk
in 1979

Promotiecommissie

Promotor	Prof. dr. G.Glas
Overige leden	Prof. dr. J. Den Boer
	Prof. dr. S. Gallagher
	Dr. J. Sleutels
	Prof. dr. M. Slors

ISBN/EAN: 978-90-9025662-7

Printed by Universal Press, Veenendaal

To all who cared without actually knowing what I was doing, and one of them in particular

Table of Contents

Acknowledgments

The Greek poet Archilochus famously wrote in one of his fables that ‘the fox knows many things, but the hedgehog knows one big thing’. This book can be seen as an attempt to cross a fox with a hedgehog, in the sense that it articulates a story about intersubjectivity by combining specific empirical findings from various scientific disciplines with a more general philosophical insight about how these findings should be interpreted and what they tell us about our everyday interactions with others. This was not an easy task, and I am very thankful for the many helping hands I received during this process.

I owe much to Gerrit Glas, who arranged a PhD position for me at the University of Leiden, where I began my work in August 2005. Gerrit has been an inspiring teacher and caring tutor, and gave me a lot of freedom to find my own path in philosophy. I am also very grateful to Shaun Gallagher and Daniel Hutto for providing me with the opportunity to spend several months at respectively the University of Central Florida in the USA and the University of Hertfordshire in Great-Britain. Although their philosophical backgrounds are quite different (Gallagher is very much rooted in the phenomenological tradition, whereas Hutto takes a more analytic approach to philosophy), this tension has been one of the main motivational forces responsible for the realization of this book. Many of the ideas I put forward are inspired by their writings and/or extracted from my conversations with them. Another key figure has been Marc Slors, who gave me a warm welcome to his Nijmegen research group, which is arguably one of the most promising philosophy of mind communities in the Netherlands. Especially my collaboration with Derek Strijbos has been very fruitful: not only did it help to structure my thinking, but we also managed to produce several good articles together. This is also true for my collaboration with Sanneke de Haan (University of Heidelberg), who I very much enjoyed working with and hope to continue doing so in the near future.

Victor Gijsbers and Wout Cornelissen (in arbitrary order) have been my closest friends at the Leiden University, and I want to thank them for the personal support they gave me and for the many hours we spent discussing and debating the ins and outs of philosophy and all the other things that make life interesting. Lies Klumper deserves a big ‘thank you’ as well. Being the center of gravity of the philosophy department, she was always ready to offer a kind word, a smile, or a good cup of tea. I’m also very grateful to my other Leiden colleagues: Jeroen van Rijen, Eric Schliesser, Bruno Verbeek, Pauline Kleingeld, Jan

Sleutels, James McAllister and all the others for their kind advice on various (philosophical) matters.

I have attended several research groups during my research over the last years, and this book has much benefited from the many helpful discussions and opportunities to present my work for various audiences. I already mentioned the Nijmegen research group lead by Marc Slors. I have also very much enjoyed my regular meetings with a small circle of psychiatrists from the Dimence Institute for Mental Health. I am particularly grateful to Ewoud de Jong, who gave me the opportunity to be involved in a very interesting EEG-experiment, and Bram Sizoo, who allowed me to sit in on his ASD diagnostic sessions. I also want to thank the participants of the Utrecht research seminar for analytic philosophy (organized by Herman Philipse and Rik Peels) for their vigorous and constructive comments on my work.

During the third year of my research, Gerrit Glas encouraged me to acquire some ‘hands on’ experience in the field of cognitive neuropsychology. I would like to thank Nelleke van Wouwe for giving me the opportunity to participate in her fMRI project, and for spending several weekends with me in the LUMC in order to collect the brain imaging data necessary for this study. My thanks also go to Henk van Steenbergen, Serge Rombouts, Bernhard Hommel, and André Keizer for their assistance with the (quite many) questions and problems I encountered during the FSL data analysis and their willingness to extensively discuss the results of this study. None of the findings made it into this book, but the scientific know how I gained during this period was priceless.

I also would like to acknowledge a number of institutions that provided me with the financial support for conference visits, workshops and summer schools. The Leiden University Fund supported several trips abroad, as well as my stay at the University of Central Florida in the USA and the University of Hertfordshire in Great-Britain. The Dutch Association for Psychiatry funded my participation in the yearly INPP conferences. The VolkswagenStiftung and the European Science Foundation provided financial support for various interdisciplinary workshops and summer schools.

There are probably many others who I have not mentioned and to whom I apologize for my lapse of memory. I am immensely grateful to Josephine Lenssen for proofreading this book and commenting on the earlier drafts. My family and friends I want to thank for their personal support and the much-needed distractions they offered from work. But the person who deserves the most credit is my wife, Carla, who has always been very supportive and generously tolerated my preoccupations and travels. She has been a constant source of inspiration.

Philosophy of mind, paradoxically enough, became an interesting area of philosophy only when philosophers began to stop taking the notion of ‘mind’ for granted and began asking whether it was a misleading locution

- Rorty 1982

Prologue

From Theory to Practice

Philosophers and psychologists tend to inscribe their own project of inquiry into our ordinary methods of understanding one another, so that in the context of everyday life we too are presented as navigating our social world primarily by observing, hypothesizing, predicting how creatures like us operate.

- McGeer 2001

The problem of intersubjectivity

This book is about what happens when two people meet. Or perhaps it is better to say that it is about what *precedes* such an encounter, since it attempts to spell out the *preconditions* of our meetings with others. It tries to capture the practices and processes that enable and facilitate these meetings in the most basic of ways, in order to lay down the 'rules of engagement'. Its main aim is to present an account of *intersubjectivity.* We sometimes use the word 'empathy' to denote the experience of similarity that arises when our encounters with others go well. This book, however, importantly goes beyond empathy insofar it rejects the idea that we can explain our face-to-face encounters in terms of a specific and particular mode of consciousness. Instead, it emphasizes that our ability to engage with others cannot be taken as a brute fact, since this ability is conditioned, structured and shaped by our bodily existence and social embeddedness.

Most contemporary explanations of intersubjectivity fall into two main categories: theory theory (TT) and simulation theory (ST). Theory theory argues that our encounters with others depend on the ability to employ a folk psychological theory (a 'theory of mind') in order to explain and predict their behavior. Some rationalist-inclined theory theorists claim that such a theory is already there from the very moment we are born - in the form of a sophisticated, inherited biological device they call a 'mindreading module’. Young children use only some of its basic principles, but over the course of development, their ability to exploit what they already know increases (cf. Fodor 1995). Other more empiricist-oriented theory theorists stress that the ability to explain and predict others’ behavior is not innate, but develops as children increasingly start to experiment and explore the world. According to this ‘child scientist’ approach, children proceed in very much the same way that scientists proceed, getting new evidence and revising their folk psychological theories in light of it (cf. Gopnik and Meltzoff 1997).

Simulation theory rejects the idea that our understanding of others requires a theory. Instead, it proposes that social encounters are primarily about putting ourselves in the others’ shoes, imagining what we would do (think, feel etc.) in their situation. According to proponents of ‘offline simulation’, such a process is driven by pretend mental states that are fed into our own offline decision-making mechanism (cf. Goldman 2006). Advocates of ‘actual simulation’, on the other hand, argue that simulation has a much more basic function: it enables us to apply what are essentially first-person decision procedures to others by transforming ourselves into other ‘first persons’ (cf. Gordon 1995).

Despite the fact that TT and ST are often portrayed as bitter rivals, they have a lot in common. A good way to get an initial feel for what drives both positions is to see them as providing an answer to a fundamental question about intersubjectivity: how are we able to recognize that other persons are ‘mind-endowed’ in the first place? John Stuart Mill (1889) formulated the question as follows: ‘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 figures which I see and hear, have sensations and thoughts, or in other words, possess Minds?’ (p.243). This is nowadays referred to as *the problem of the other mind*. Mill also offered a possible solution to this problem: the argument from analogy. He argued that, since I know my own mind and how it relates to my body, I am able to infer that this is probably also true for other persons on the basis of an *analogy* between our bodies.

The argument from analogy is still the point of departure for most versions of ST. However, it can be objected that Mill’s solution is flawed, because it represents one’s knowledge of the other mind as resting on an inductive generalization from exactly *one* case. Therefore, TT tackles the problem from a rather different angle. It claims that mental states such as beliefs and desires are *theoretical unobservables*, and maintains that we are justified in postulating them as long as this yields an appropriate amount of explanatory and predictive power (cf. Churchland 1988). I will explore and discuss the specifics of both TT and ST, and the way they deal with the problem of the other mind extensively in the coming chapters.

It is important to realize that in their attempts to come up with an answer to the problem of the other mind, both TT and ST agree with many of its undergirding assumptions. Moreover, these assumptions are arguably a decisive source of inspiration for their take on intersubjectivity. In what follows, I will highlight the ones that are particularly important to our discussion:

(i) In the first place, there is the idea that our encounters with others are intrinsically *problematic*. The problem of the other mind suggests that *doubt* is at the heart of intersubjectivity: how can we be sure about the existence of the other mind? TT and ST follow in Mill's footsteps by depicting our everyday encounters with others as complicated puzzles, uncertain expeditions to a remote and unknown region called 'the other mind' with the primary objective of gaining knowledge of what goes on there.

(ii) To accept the problem of the other mind as a genuine problem is not only to conceive of social interaction as a quest for certainty, however. It is also to accept a certain conception of the mind it brings along. TT and ST interpret the mind as an isolated 'I' - an autonomous entity that represents the outside world and its own body but is at the same time separated from it. They conceive of the mind as a mysterious inner realm, hidden away behind the overt behavior we can see. This conception of the mind has a rich historical background, and in their attempts to trace its origin, philosophers are often quick to point the finger at Descartes, the godfather of modern philosophy of mind. Such accusations are certainly not unfounded. At the same time, however, we should take into account that for Descartes, the existence of the other mind was not yet problematic – he was able to evade the solipsistic consequences of his method of doubt by appealing to a benevolent God. But the specter of solipsism started to loom ever more threateningly in the works of Descartes' successors, particularly in those of the British empiricist tradition who no longer accepted such a theological appeal. It is therefore scarcely surprising that, for a philosopher such as Mill, the problem of the other mind becomes an 'official' problem.

(iii) Another important assumption is that our doubts about the other mind can be overcome by a self-conscious, methodological and critical way of thinking. Descartes thought that a strict introspective method was the only road to certain knowledge, since it provided the user with an immediate awareness of the mind's ideas. These ideas, supported by divine authority, also guaranteed the existence of the other mind. Mill, on the other hand, like his contemporaries, no longer wished to invoke God to assure him of the existence of the other mind, and therefore sought its justification in radically different terms. His argument from analogy postulates an *inferential process* that enables us to come up with empirical generalizations between our mental states and our bodily behavior, which then in a further step can be attributed to other people. However, despite the huge differences between Descartes and Mill, both thought that intersubjectivity depended on a conscious, cognitive process - a stepwise procedure initiated by a hyper-reflexive agent. And this idea is very much alive in contemporary articulations of TT and ST. It is telling that intersubjectivity is nowadays often understood in terms of 'folk psychology', a label used to emphasize that our common-sense understanding of others is actually nothing more than a folk-version of the methodology employed in the science of psychology.

(iv) Last but not least, it is generally assumed that thinking or cognition functions as an *intermediary* between perception and action. Hurley (2008) calls this the 'sandwich model' of intersubjectivity, since it regards 'perception as input from the world to the mind, action as output from the mind to the world, and cognition as sandwiched in between' (p.2). According to the sandwich model, our meetings with other minds are structured in the following way: we start out by observing another agent's bodily behavior, but at this point, we don't yet have evidence for the existence of his mind or any clue about the mental states he is currently entertaining. In order to get there, we need to engage in an inferential and/or deliberative process. When this process is brought to a satisfying conclusion, we are ready for (inter)action. It goes too far to trace the historical roots of the sandwich model here. For now, it is sufficient to point out that both Descartes and Mill were each in their own way committed to this model. So are many contemporary versions of TT and ST.

The practice of mind

This book, in one clear sense, seeks to undermine the picture of intersubjectivity sketched above and the various problems that result from it. But it does not want to do so by simply denying its underlying intuitions. Instead, it aims at a more constructive approach by showing what intersubjectivity looks like from a *pragmatic* point of view. Most explanations of intersubjectivity that stress the importance of theory, such as TT and ST, end up modeling our knowledge of the other mind on the perceptual abilities of the *individual agent*. This inevitably leads to what Dewey (1960) called a ‘spectator’ theory of knowledge. The pragmatic perspective I want to promote, however, emphasizes the *interactive* instead of perceptual nature of our knowledge of the other mind. The word ‘pragmatic’ is derived from the Greek word ‘pragma’, which means ‘action’. However, it also lies at the basis of the word ‘practice’. The centre of gravity of this book is the idea that intersubjectivity is enabled through a large variety of *second-person practices*. These practices structure our encounters with other minds and provide us with the social tools needed to understand them. Thus, we might say that the primary focus of this book is on the *practice of mind*.

My pragmatic account of intersubjectivity does not so much elaborate on one single theory, but rather unites and integrates a number of recent insights and proposals that have been made with regard to social interaction. It borrows from *enactivism* insofar as it endorses the aphorism: ‘knowing is doing is being.’ Neither our being in this world, nor our knowledge of it is pre-existent, in the sense that it is given beforehand. Instead, it is *enacted*, arising from our moment-to-moment coping with the environment and other people. The adaptive process wherein identity and knowledge are constantly emerging as the result of interactions with the environment is what we call learning: a continuing exploration of an ever-evolving landscape of possibilities and of selecting (not necessarily consciously) those actions that are adequate to maintain one’s balance.

This has important consequences for our conception of identity and knowledge, of ‘mind’ and ‘world’. Enactivism rightly emphasizes that the mind is fundamentally shaped by its bodily existence (embodiment) and cannot be understood in isolation from its environment (embedment). The focus of this book is in particular on how this mind-shaping has to be understood in relation to our interactions with *other minds*. Pursuing an enactivist agenda also has important consequences for our conception of the world - our knowledge of the environment. According to enactivism, knowledge is not ‘the representation of a pre-given world by a pre-given mind’, but it is rather ‘the enactment of a world and a mind on the basis of a history of the variety of actions that a being in the world performs’ (Varela et al. 1991, p.9). This book tries to explain how this process of enactment provides us with knowledge of the other mind.

Besides its obvious affinities with enactivism, my pragmatic proposal builds on the insights of several philosophers from both the phenomenological and the analytical tradition. It draws on the phenomenological tradition (the work of Shaun Gallagher in particular) in order to question the phenomenology of uncertainty that is presupposed by TT and ST, and argue that what is at the core of our everyday social encounters is not exclusively a knowledge-affair. On the contrary, much of what goes on during these face-to-face meetings actually happens *before we know it*. Moreover, the phenomenology of everyday intersubjectivity suggests that the explicit kind of meta-cognitive theorizing presupposed by many versions of TT and ST 'is not our everyday practice; it is not the way we think of ourselves or of others' (Gallagher 2004, p.202). My proposal draws on the analytical tradition insofar it follows philosophers such as Ludwig Wittgenstein and Wilfred Sellars (and their contemporary representatives such as Daniel Hutto) in their view of the relation between language, mind and meaning. Most importantly, I use their insights to criticize the attempt to model our knowledge of the other mind on a first-person 'immediate awareness' of one's own mind, as ST does, or on a third-person theoretical understanding of psychological principles, as TT does. The lesson I take from them is that neither what we call 'mind', nor 'world', is *presupposed by* or *constitutive for* social interaction. Rather, both *emerge from* the linguistic practices that structure second-person interactions. Therefore, instead of appealing to a private language or a set of implicit theoretical principles, the pragmatic approach to intersubjectivity I have in mind pays attention to *actual linguistic practices* since these make it possible for us to deploy such vocabularies in the first place.

An important aim of this book is to stretch intersubjectivity beyond the limits of 'folk psychology', or what has recently become its substitute term: 'mindreading'. This is not to say that mindreading does not play any role in our encounters with others. But on my proposal, its role is relatively modest and its function different from what is generally assumed. The consensus has it that mindreading is primarily about the generation of reliable predictions and explanations of others' actions. It is often assumed that this depends on a very basic (innate) capacity that is mainly exercised in third-person theoretical contexts - situations in which the interpreter is a bystander, someone observing the agent performing the action without interacting with him. This book, however, presents a view of mindreading as firmly rooted in a rather advanced, second-person practice, and also promotes a very different picture of reason explanation. It takes to heart Hutto's (2004) advice that 'taking seriously the second-personal starting point ought to provoke us to reconsider the [...] prevailing views about the function and context of much commonsense psychology, even when it comes to its most characteristic activity of providing reason explanations. In abandoning the idea that the contexts in which we make sense of others are normally spectatorial, we can recast and re-orient our thinking about the nature of our expectations about each other and about how such explanations are ordinarily achieved' (p.550).

Pragmatism and its limits

Restricting the scope of folk psychology allows for an explanation of intersubjectivity that goes above and beyond those of a purely 'mentalistic' variety, and paves the way for an appeal to evidence that is *interdisciplinary* in nature. This book draws on various disciplines, such as experimental psychology, neuroscience, studies of pathology and developmental psychology, and uses their findings to support the large range of practices it puts forwards. For example, many anticipatory and predictive processes that facilitate our meetings with others are dependent on low-level sensorimotor processes that can be described in terms of neurobiological mechanisms. To a certain extent, these processes allow for 'hands-free' intersubjectivity and can be used to explain why a large part of our encounters with others does not require conscious reflection at all.

This naturally leads to questions about the status of empirical evidence in the debate on intersubjectivity. Although the brand of pragmatism I want to articulate pays a lot of attention to scientific findings, this does not mean that it advertises reductionism or instrumentalism. Nor does it wish to promote a kind of scientism. Rather, it starts by taking intersubjectivity at face value and closely studies what people are actually *doing* when they are trying to understand others and what happens during these encounters. The kind of pragmatism I have in mind focuses on *actual second person practices*. It asks: How can we describe what is going on? And: How does it come about? The first question addresses the phenomenology of intersubjectivity, and to answer it properly we require something along the lines of what Gallagher (2006) calls 'front-loaded phenomenology': a good description of the way we experience our everyday encounters with others, which can then be used as input for scientific experimentation. The second question suggests that we can tackle many problems pertaining to the various elements of intersubjectivity by investigating how intersubjectivity comes about, that is, by identifying its preconditions. In my opinion, the most promising approaches to intersubjectivity therefore have to engage with either its ontogenetic development or its phylogenetic evolution. The aim of this book is to do the *former* - it provides the reader with a developmental story about intersubjectivity. A short but plausible story about the evolutionary roots of intersubjectivity can be found in Hutto (2007a), who also deals with TT and ST claims on this subject. For more elaborated accounts, see for example the works of Donald (1991, 2001) or Tomasello (1999, 2003, 2008).

The pragmatic attitude towards intersubjectivity which is advocated throughout this book has not only important consequences for the way I want to approach the problem of the other mind. It also affects how I see many of the ontological problems that have traditionally set the agenda of philosophy of mind: the relationship between mind and body, mental causation, emergence, dualism, physicalism etcetera. I am convinced that these problems would benefit from a pragmatic treatment as well, but unfortunately this falls beyond the scope of this book. However, since they often linger in the background of the debate about intersubjectivity, I will occasionally bring them to the fore in order to show what they would look like through pragmatic spectacles.

Of course I realize that pragmatism, as a philosophical program, has its limits. The kind of pragmatism that I want to put forward here, however, is actually very modest. It continues and deepens a line of thought initiated by Goldman (1989), who remarked that 'no account of interpretation can be philosophically helpful [...] if it is incompatible with a correct account *of what people actually do* when they interpret others' (p.162, italics added). In other words, this pragmatism emphasizes that we cannot explain intersubjectivity without paying attention to the fact that it is something that is 'happening' between people, something that is 'done'. Its main message is: preach what you practice!

A survey of the book

The outline of this book is as follows. The first two chapters deal with what I call the 'internal' problems of TT and ST, in other words, with the problems that start to appear when one accepts a certain picture of intersubjectivity. I will advance conceptual as well as phenomenological arguments in order to show that both TT and ST offer an extremely impoverished and problematic account of intersubjectivity. These chapters also involve a critical assessment of the scientific evidence that both parties have brought forward in order to support their claims, ranging from the field of developmental psychology (e.g., results on false-belief tasks) to the realms of neurobiology (e.g., findings on mirror neurons).

It is also possible to question TT and ST approaches to social interaction at a more basic level. Such a more hermeneutically-oriented analysis allows us to uncover their deeper motivations and investigate the extent to which both are inspired by similar assumptions about intersubjectivity. These assumptions will be discussed and challenged in chapter 3.

The pragmatic view I want to propose has its starting point in the idea of intersubjectivity as building on a set of second-person practices. It further articulates and extends Gallagher's proposal (e.g., Gallagher 2005) that a wide range of *embodied practices* allow us to employ various innate or early developing capacities that provide a basic form of social understanding - what Trevarthen (1979) called 'primary intersubjectivity'. Throughout development, these capacities become more and more embedded in a broader social and pragmatic context, thereby enabling us to engage in *embedded practices* of joint attention (so-called 'secondary intersubjectivity'). This is the topic of chapter 4. Embodied and embedded practices are not self-sufficient. They depend on and are shaped by our bodily existence, and build upon the kinds of experiences that result from having a body with various sensory-motor capacities. Chapter 4 also offers an explanation of how complicated processes at the neurobiological level provide us with a minimal form of self-awareness (including a sense of ownership and agency), and a basic awareness of others (what I call 'co-consciousness').

While embodied and embedded practices constitute the base-line for social understanding and continue to do this after the development of more advanced abilities, they by no means exhaust the possibilities for intersubjectivity. *Narrative practice* comes into play with the emergence of linguistic abilities and a number of other ontogenetic achievements (such as the capacity for temporal integration, (auto)biographical memory and perspective taking), and they allow us to further fine-tune and sophisticate our understanding of self and other (Hutto 2007, Gallagher and Hutto 2008). This will be discussed in the first part of chapter 5.

Narrative practice may also explain how we enter what Sellars termed the normative 'space of reasons', and acquire the ability to make sense of actions in terms of *reasons*. The second part of chapter 5 discusses the strengths and weaknesses of Hutto's (2007) ‘narrative practice hypothesis’, according to which children come to master the art of folk psychology through direct encounters with folk psychological narratives - stories about reasons for acting. I propose that, initially, children are only capable of interpreting others’ actions in terms of reasons against the background of a *shared* world. But the acquisition of mental concepts eventually enables them to vastly expand and improve their interpretation abilities by opening up new ways of *individuating* the reasons of other agents, in a way that is tailored to their psychological make-up

1.

Theory Theory

Science is continuous with common sense, and the ways in which the scientist seeks to explain empirical phenomena are refinements of the ways in which plain men, however crudely and schematically, have attempted to understand their environment and their fellow men since the dawn of intelligence.

- Sellars 1963

Mindreading

Our everyday meetings with other minds often seem to carry with them an enormous potential for confusion and misunderstanding. Consider the following example by Pinker (1994, p.80):

First guy: I didn't sleep with my wife before we were married, did you?
Second guy: I don't know. What was her maiden name?

Yet, for the most part, our social engagements proceed smoothly. Mistakes such as in the above example are the exception rather than the rule. In fact, at a second glance we might even wonder whether the example presents a case of genuine *misunderstanding*. It is obviously not the intention of the first speaker to suggest that both he and the second speaker might have slept with the same woman. In overhearing this exchange, most of us would probably assume that the second speaker fully understands what the first speaker is driving at, but chooses to ignore the intention behind the question in order to make a *joke* of it. Normally, we do not only pay attention to the actual words a speaker uses. When a cop shouts 'Drop it!' a robber is usually not left in a state of acute doubt over the ambiguity of the term 'it'. On the contrary, he immediately realizes that the word ‘it’ refers to the gun in his hand. But how is he able to do this?

According to contemporary explanations of intersubjectivity, this requires a considerable amount of *mindreading.* The idea is that by engaging in some kind of special cognitive procedure, we are able to discover and specify the mental states of others and use them in order to explain and predict their actions. This often implies that we have to decode their actual speech, and go away beyond the words we hear to hypothesize about their possible intentions. Baron-Cohen (1995) argues that this is exactly what happens in the ‘drop-it’ example: 'the robber makes the rapid assumption that the cop meant (i.e., intended the robber to understand) that the word “it” should refer to the gun in the robber's hand. And at an even more implicit level, the robber rapidly assumes that the cop intended to recognize his intention to use the word in this way’ (p.27). This kind of mindreading is thought to be of central importance to the logic of everyday sense-making, no matter whether it concerns verbal or non-verbal communication. It is fundamental to our intersubjective understanding. Nichols and Stich (2003) put it like this: ‘[...] we engage in mindreading for mundane chores, like trying to figure out what the baby wants, what your peers believe about your work, and what your spouse will do if you arrive home late’ (pp.1-2).

Consider another example from Pinker (1994, p.227):

Woman: I'm leaving you.
Man: Who is he?

Although it is sometimes said that men are lacking in the communication department, this man seems to need only a few words to figure out what is going on. Baron-Cohen (1995) claims it is again mindreading that does the trick here. In order to come up with this phrase, the man ‘must have thought [formed a belief] that the woman was leaving him for another man’ (p.28). Moreover, Baron-Cohen also suggests that *we ourselves* (when overhearing this exchange) must attribute this belief to the man in order to make sense of the conversation. Otherwise, the dialogue would seem ‘disconnected, almost a random string of words’ (ibid.). Our mindreading is able to fill in the ‘gaps’ in communication and ‘holds the dialogue together’ by representing the mental states that could have been in the man’s mind. In other words, mindreading is a must-have because without it, we are simply unable to make sense of others. The attribution of mental states to others is our natural way of understanding the social environment. In the words of Sperber (1993), ‘attribution of mental states is to humans as echolocation is to the bat’. Without mindreading, the other mind remains a mystery.

When it comes to explaining the ins and outs of mindreading, philosophers typically (and often exclusively) focus on the mental states of *belief* and *desire*.[1] Russell (1940) called these mental states propositional attitudes, since they are psychological attitudes that exhibit a special kind of intentionality - an ‘aboutness’ or directedness toward possible situations.[2] A belief is usually defined as a cognitive attitude that aims at truly representing how things stand with the world, whereas a desire is defined as a motivational attitude that specifies a goal for action. What is so attractive about mindreading is that it allows us to exploit specific combinations of these beliefs and desires for the purposes of both behavior explanation and prediction. In case of behavior prediction, we start with two interlocking beliefs and desires and work our way towards a predicted or anticipated behavioral outcome, whereas in case of behavior explanation, we work back from the behavior under consideration to a particular belief-desire pair. Mindreading, thus understood, is not only thought to be the *primary* but also the *universal* mode of intersubjectivity. Fodor (1987), for example, remarks that: ‘There is, so far as I know no human group that doesn’t explain behavior by imputing beliefs and desires to behavior (And if an anthropologist claimed to have found such a group, I wouldn’t believe him)’ (p.132).

Over the last decades the importance of mindreading for intersubjectivity has been promoted by two main approaches: theory theory (TT) and simulation theory (ST). In this chapter I am concerned primarily with *theory theory*. First, I briefly introduce the historical background of TT in order to shed light on some of its basic assumptions (section 1). I then

[1] The assumption that mindreading is rooted in belief-desire psychology is taken for granted by almost all participants in the intersubjectivity debate. Currie and Sterelny (2000), for example, assert that ‘our basic grip on the social world depends on our being able to see our fellows as motivated by beliefs and desires we sometimes share and sometimes do not not [...] social understanding is deeply and almost exclusively mentalistic’ (p.143). And Frith and Happé (1999) state that ‘in everyday life we make sense of each other’s behavior by appeal to a belief-desire psychology’ (p.2).

[2] Propositional attitudes are relational mental states that connect a person to a proposition. They are often assumed to be the simplest components of thought and can express meanings or contents that can be true or false. In being a type of attitude they imply that a person can have different mental ‘postures’ towards a proposition, for example, believing, desiring, or hoping, and thus they imply intentionality.

proceed to discuss the various TT positions in further detail, touching on a number of problematic issues along the way (section 2 and 3). Next, I review the empirical evidence that is frequently put forward in support of TT, and raise some questions with regard to its interpretation (section 4). In the final part of this chapter, I address the problem of eliminativism and present a concise summary of TT-related problems (section 5). Together, these problems cast some initial doubt on TT explanations of intersubjectivity.

1.1 Folk psychology as theory

According to the TT approach to intersubjectivity, the ground rules for mindreading are laid down by what is generally referred to as *folk psychology.* [3] In spite of its commonsensical (or intuitive) nature, folk psychology is essentially a *theory,* which explanatory and predictive virtues are what make mindreading such a powerful tool in understanding others. Churchland (1986) describes folk psychology as the 'rough-hewn set of concepts, generalizations, and rules of thumb we all standardly use in explaining and predicting human behavior. Folk psychology is commonsense psychology - the psychological lore in virtue of which we explain behavior as the outcome of beliefs, desires, perceptions, expectations, goals, sensations and so forth. It is a theory whose generalizations connect mental states to other mental states, to perceptions, and to actions. These homey generalizations are what provide the characterization of the mental states and processes referred to; they are what delimit the 'facts' of mental life and define the explananda' (p.299).

The basic idea behind TT is that the folk psychological knowledge that fuels our mindreading skills is continuous with scientific knowledge. The latter is a more methodical, systematic, and controlled version of the former, but the two are fundamentally alike in the

[3] There is a lot of confusion about the notions of mindreading and folk psychology, since they are often used interchangeably. On top of that, the label folk psychology itself is somewhat unfortunate because it tends (and was intended) to invoke a comparison between our commonsensical understanding of others and the scientific explanations of behavior in psychology. In this book, the term mindreading is generally used in a broad sense, referring to the ability to interpret others in terms of mental states such as beliefs and desires, whereas the term folk psychology is used to denote the more specific (TT) idea that this ability has a theoretical basis. But this distinction is a bit artificial, since it is questionable whether we can make sense of mindreading without any appeal to theory whatsoever (cf. chapter 2.1).

sense that both are thoroughly *theoretical* and *fallible*. A good starting point to understand the consequences of this idea is the work of Wilfred Sellars, in particular his criticism of the so-called 'myth of the given' (cf. Sellars 1963). Sellars was fervently opposed to the empiricist claim that scientific knowledge has a foundation because some of our claims about the world have a privileged epistemological status, in the sense that they are 'given' to us in our first-person experience.[4] One of the main objectives of empiricism had been to prove that observational knowledge could 'stand on its own feet', and this was precisely what Sellars denied. He remarked that 'the idea that epistemic facts can be analyzed without remainder - even "in principle" - into non-epistemic facts, whether phenomenal or behavioral, public or private, with no matter how lavish a sprinkling of subjunctives and hypotheticals is [...] a radical mistake - a mistake of a piece with the so-called "naturalistic fallacy" in ethics' (p.131). Science is rational, according to Sellars, not because it has a foundation in our first-person experience of 'sense data' (the content of one's perceptual experience), but because it is a social, self-correcting enterprise 'which can put any claim in jeopardy, though not all at once' (p.170).

To counter the myth of the given, Sellars constructed his own piece of 'anthropological science fiction', in which he speculated that our private vocabulary (the folk psychological terms we use to describe our inner life) might have originally been *postulated* rather than *observed*. The 'myth of Jones' tells us how our fictive Rylean ancestors, who were only familiar with some sort of methodological behaviorism, might have come to develop a non-observationally based understanding of such vocabulary. This revolution in social understanding is attributed to a genius called Jones, who discovers that by modeling the 'inner episodes of thought' of his companions on their overt speech acts, he is able to explain and predict their future behavior, even in the absence of verbal reports. In a later stage of development, Jones and the others also learn to apply the 'theory' to themselves: 'Once our fictitious ancestor, Jones, has developed the theory that overt verbal behavior is the expression of thoughts, and taught his compatriots to make use of the theory in

[4] One class of these 'givens' that has traditionally been privileged concerns the claims about one's own 'sense data', or the contents of one's perceptual experience. Their special epistemological status is backed up by the following argument: my sincere claim that I see a red object might well turn out to be mistaken, but my claim that I am now experiencing red sense data – 'as if' I were seeing a red object – could not possibly turn out to be mistaken. Another class of privileged claims contains those claims that concern one's apparent memories and beliefs. I can't be certain that I have indeed seen a red object, but I certainly seem to remember seeing one – and although the belief that I have seen one might be false, the sincere claim that I believe so cannot be mistaken.

interpreting each other's behavior, it is but a short step to the use of this language in self-description [...] Our ancestors begin to speak of the privileged access each of us has to his own thoughts. What began as a language with a purely theoretical use has gained a reporting role' (p.320).

Sellars' account of the origins of our folk psychological vocabulary has undoubtedly been a great source of inspiration for the TT picture of mindreading as being essentially *theory-driven*.[5] Most importantly, this is because Jones is portrayed as a first-rate scientist, who constructs his model of non-observational mental states in a way similar to how modern science constructs theoretical posits, and then uses it as an explanatory theory in order to make sense of the observable behavior of others. But TT also adopts the idea that the knowledge we use to mindread others is intrinsically fallible and always up for revision. Each of our beliefs about the other mind is no more than a hypothesis, and no matter how spontaneous, non-inferential or intuitively evident it might seem, it remains a conjecture that can in due course come to be revised. Unfortunately, this is also true for the ensemble of law-like generalizations, rules of thumb and interconnected concepts we call folk psychology, which raises the worry that folk psychology *as a theory* might turn out to be a 'false and radically misleading conception of the causes of human behavior and the nature of cognitive activity' (Churchland 1988, p.43). This is a serious problem for proponents of TT who take mindreading to be the primary mode of intersubjectivity. Fodor (1987), for example, remarks that if the ordinary person's understanding of the mind should turn out to be seriously mistaken, it would be 'the greatest intellectual catastrophe in the history of our species' (p.xii). This possibility is further explored in the last section of this chapter.

Sellars' claim that knowledge is thoroughly fallible, a theme also developed by Quine (1953) and Feyerabend (1962), really starts to hurt when we realize that it not only applies to our knowledge of others, but also has implications for our *self-knowledge*. The bottom line of the myth of Jones is that privileged access and the articulation of a private vocabulary do not come *first,* but rather are derivative, secondary capacities that depend on a more basic language with 'a purely theoretical use'. As a result, the knowledge I have of my own mind can no longer serve as a reliable springboard for the acquisition of knowledge of the other mind. According to the argument from analogy, we can infer that the bodily behavior of others is probably linked up with a mind because we are already

[5] Bermudez (2003), for example, notices that the idea of folk psychology as an explanatory theory is 'much to the fore [...] in Sellars' influential mythical account of how folk psychology might have emerged' (p.47).

endowed with an intimate knowledge of how this works in our own case. But most versions of TT follow Sellars' suggestion that we cannot just assume that self-knowledge is 'given', and therefore argue that both self and other knowledge are *equally* problematic. It is only because mindreading is driven by a folk psychological theory that we are able to make sense of others and ourselves in the first place. With these general comments in mind, let us now take a closer look at the various flavors of TT.

1.2 A taste of TT

TT explanations of intersubjectivity can be divided into two broad categories: internalist and externalist versions (cf. Stich and Ravenscroft 1994). The internalist division of TT claims that our mindreading abilities depend on an internal 'theory of mind'. But even within this camp there are different stories about how we acquire such a theory and how it enables us to read the mental states of others.

The 'modular' subdivision of internalist TT argues that our theory of mind is based on an innately specified, domain specific mechanism (Fodor 1983, Leslie 1991, Baron-Cohen 1995). This view is mainly inspired by Noam Chomsky (1957), who speculated about the existence of a universal, generative grammar grounded in an underlying language acquisition device - a dedicated and autonomous brain module for the rapid learning of language. In a similar vein, modular TT (or MTT) claims that there has to exist some kind of 'mindreading module', a sophisticated biological device that contains all the ingredients for a universal folk psychological theory. Tooby and Cosmides (1995), for example, argue that 'humans everywhere interpret the behavior of others in […] mentalistic terms because we all come equipped with a "theory of mind" module [...] that is compelled to interpret others this way, with mentalistic terms as its natural language' (p.xvii). When it comes to the ontogenetic development of such a mindreading module, some advocates of MTT have suggested that it is in place from the moment of birth, such that 'the child's theory of mind undergoes no alteration; what changes is only his ability to exploit what he knows' (Fodor 1995, p.110). Accordingly, young children use only some of the theoretical principles contained in the module, effectively operating with a very simple theory of mind. Many theory theorists see the existence of an innate theoretical module as a biological endowment, a gift from our evolutionary ancestors that allows for a rapid explanation and prediction of another organism's behavior (cf. Baron-Cohen 1995). This view is often complemented by the 'Machiavellian intelligence' hypothesis, according to which a primary selection pressure driving human brain development was strategic interaction, with social competition leading to increasingly sophisticated mindreading mechanisms (e.g. Byrne and Whiten 1988).

There are also versions of MTT that are committed to a less substantial innate component. For example, Garfield et al. (2000) claim that mindreading is supported by an 'acquired module', which forms through the interaction between innate capacities and social environment, thus emphasizing the importance of developmental processes. And scientific TT (or STT) downplays the importance of an innate module even further. It claims that, with the exception of a number of specific theoretical principles, our theory of mind is not innate but acquired through a course of development: children develop their everyday knowledge of the social world by using the same cognitive devices used in science. They proceed like little scientists, testing and revising their hypotheses about other minds in the light of new evidence (Gopnik and Wellman 1992, 1994; Gopnik and Metzoff 1997). Therefore, STT is also nicknamed 'the child-scientist hypothesis'.

Innateness and the problem of learning

According to Alison Gopnik (2003), the main difference between STT and MTT can be traced back to the age-old rationalist/empiricist dispute about the problem of knowledge: the question of how to overcome the unbridgeable gap between our abstract complex, highly structured knowledge of the world, and the concrete, limited and confused information provided by our senses. The rationalist way to solve this problem, Gopnik argues, is to realize that although it looks as if we learn about the world from our experience, we don't really. Actually, we knew about it all along. The most important things we know were there to begin with, 'planted innately in our minds by God or evolution or chance' (2003, p.238). The empiricist, on the other hand, claims that although it looks as if our knowledge is far removed from our experience, it isn't really. If we rearrange the elements of our experience in particular ways, by associating ideas, or putting together stimuli and responses, we'll end up with our knowledge of the world. This leads to an interesting dilemma between rationalism and empiricism. The former is very well able to account for the abstract, complex nature of knowledge, but cannot explain, and therefore denies, the fact that we learn. The latter is able to explain learning, but can’t explain, and so denies, the fact that our knowledge is so far removed from experience.

Gopnik proposes that STT should be seen as the *empiricist* reaction to the *rationalist* line of thinking about the problem of knowledge laid down by Chomsky. Chomsky offered a particular rationalist hypothesis, the so-called ‘innateness hypothesis’ as an empirical answer to the problem of knowledge. But, as Gopnik points out, his arguments for doing so did not follow from empirical studies on the development of language and thought in children. On the contrary: ‘Chomsky’s most important argument for rationalism is the same argument that Socrates originally formulated in the Meno, it has come to be called the poverty of the stimulus argument. The learning mechanisms we know about are too weak to derive the kind of knowledge we have from the kinds of information we get from the outside world’ (2003, p.239). What Gopnik seems to suggest here is that Chomsky’s innateness hypothesis is only appealing as long as we lack real insight and understanding of our learning mechanisms. This indeed makes sense when we consider one of the main champions of MTT, Jerry Fodor. In ‘The Language of Thought’ (1975), Fodor argues that we simply *have* to accept the idea that the mind is endowed with many complex (mental) concepts prior to its arrival in this world, since only such an ‘extreme innatism’ can explain how we acquire them. The appeal to innateness is unavoidable because we lack a decent story about concept acquisition.

Gopnik, by contrast, argues that a proper empiricist solution to the problem of (folk psychological) knowledge has to avoid an appeal to innateness. Instead, it should stress the *plasticity* of learning mechanisms. If we define a theory as a learning mechanism that assigns representations to its inputs and employs a set of rules to operate on them, we should be open to the idea that the resulting representational patterns might in turn be able to *alter* the very nature of the relations between these inputs and representations. New inputs generate new representations, and in this way the very rules that connect inputs and representations can change as well. Eventually, according to Gopnik, we may end up with a system that not only has a completely renewed stock of representations, but also works with a totally different set of relations between inputs and representations than the system we started out with. She invokes Neurath’s philosophical metaphor to illustrate that STT sees knowledge as a boat that we perpetually rebuild as we sail in it. ‘At each point in our journey there may be only a limited and constrained set of alterations we can make to the boat to keep it seaworthy. In the end, however, we may end up with not a single plank or rivet from the original structure, and the process may go on indefinitely' (2003, p.242).

Theory change/evolution is possible because theories themselves build on, revise or replace earlier theories. But where do these earlier theories come from? Gopnik thinks that the answer to this question is simple: 'They are the theories we are, literally, born with. We learn by modifying, revising and eventually replacing those earlier theories with later ones' (p.244). But this prompts another question. What about the ambition to offer an empiricist alternative to the innateness hypothesis without appealing to innateness? Gopnik holds that the kind of theoretical innateness that is presupposed by STT is importantly different from Chomskyan innateness, since the former claims that the basic theories we start out with are immediately subject to radical and continuing revision in the light of the further evidence we accumulate in the course of development. But this is clearly not sufficient to conceal the fact that STT owes much of its credibility to the assumption that these innate theories indeed exist. In fact, its disagreement with MTT seems to be not so much about the innateness of folk psychological *rules*, but rather about the innateness of folk psychological *content*.

Another challenge for STT is to explain how it is possible that all children eventually come up with the *same* folk psychological theory. Goldman (1989) formulates the problem as follows: 'Another possible mode of acquisition is private construction. Each child constructs the generalizations for herself, perhaps taking clues from verbal explanations of behavior. But if this construction is supposed to occur along the lines of familiar modes of scientific theory construction, some anomalous things must take place. For one thing, all children miraculously construct the same nomological principles. This is what the (folk-) TT ostensibly implies, since it imputes a single folk psychology to everyone. In normal cases of hypothesis construction, however, different scientists come up with different theories' (pp.167-8).

Belief-desire psychology and the problem of context-sensitivity

Although it is often suggested that folk psychology includes much more than the ability to make sense of others in terms of beliefs and desires, there is a strong consensus that it should at the very least include this ability.[6] Since philosophical orthodoxy has it that individual beliefs cannot cause actions on their own, and lone desires are aimless without guiding beliefs, it is thought that we need to discover a proper *combination* of them in order to understand others and predict or explain their actions.

Both modular and scientific theory theorists agree that the folk psychological rules by which we pick out these belief-desire combinations form the core of our theory of mind. Gopnik and Meltzoff (1997), for example, claim that the theory '[...] has many complexities but also a few basic causal tenets [...] These tenets are perhaps best summarized by the "practical syllogism": if a psychological agent wants event y and believes that action x will cause event y, he will do x' (p.126). Of course, we need more than a simple practical syllogism in order to select the specific *contents* of the beliefs and desires over which the theory quantifies in a particular situation. According to most theory theorists, this requires additional theory about how beliefs and desires relate to perceptions, bodily expressions, (verbal) behavior and other mental states. Although some of these auxiliary folk psychological generalizations can be made explicit, it is usually assumed that they are largely stored and drawn upon *tacitly*. Importantly, these generalizations crucially depend for their accuracy on ceteris paribus clauses.[7] To be of any practical use, it is therefore vital that our mindreading takes into account the particular *context* of action. There may be other mental states to be derived from (or 'read off') behavioral evidence and environmental cues - situational factors, character traits, personal histories and behavioral limitations that exceed these clauses and make our folk psychological generalization less adequate.

The context requirement becomes problematic, however, when we realize that our folk psychological theory only consists of 'general theoretical knowledge - that is the sort of non-content specific knowledge that might very plausibly be held to be innately given' (Carruthers 1996, p.24). For mindreading to be *structurally* successful, folk psychological generalizations should be embedded in extensive know-how concerning their context-

[6] Hutto (2007), for example, claims that 'At a bare minimum, folk psychology *stricto sensu* is belief/desire propositional attitude psychology' (p.115, italics in original).

[7] Horgan and Woodward (1985) stress the importance of this 'all else being equal' in belief-desire reasoning as follows: 'if someone desires that p, and this desire is not overridden by other desires, and he believes that an action of kind K will bring it about that p, and he believes that such an action is within his power, and he does not believe that some other kind of action is within his power and is a preferable way to bring it about that p, then *ceteris paribus*, the desire and the beliefs will cause him to perform an action of kind K' (p.197).

sensitive application. But if we stay within the framework of TT, it seems that this know-how should itself be governed by yet another layer of tacit knowledge of rules specifying the conditions for their application. This is how Shaun Gallagher (2004) puts it: ‘We are led to ask, then, how we obtain the necessary background knowledge about others and about the various pragmatic contexts in which we encounter them. Because gaining this knowledge already involves some understanding of others, either we already have an innate theory of mind that enables this understanding, or we have some other pretheoretical, preconceptual access to others. The idea that we would need a theory of mind to gain the background knowledge necessary to get a theory of mind does not necessarily involve a vicious circle, but it certainly does involve a serious hermeneutical circle, and it requires an explanation of how the process gets off the ground’ (p.203).

Even if the *plasticity* of theory formation is heavily emphasized, as in STT, it still seems hard to reconcile the simplicity of belief-desire syllogisms with the stubborn complexity of our everyday social encounters. Our understanding of others requires a ‘massively hermeneutic’ background (Bruner and Kalmar 1998) and a theory just seems to be too far removed from practice to deliver this. An appeal to innateness seems to be the only way to deal with the lack of context-sensitivity, but I agree with Gopnik that this would be nothing more than an excuse for a lack of real understanding.

Folk psychological principles ‘ain’t in the head’

Whereas both modular and scientific TT agree that the folk psychological rules that guide our meetings with others mind are innately acquired, *externalist* versions of TT argue that these theoretical principles cannot be modeled on the individual agent, since they ‘ain’t in the head’ (cf. Stich and Ravenscroft 1994). Instead, they systematize the folk psychological ‘platitudes’ that people readily recognize and assent to - generalizations that are ‘common knowledge’ amongst ordinary folk.

Some philosophers have argued that these generalizations might be usefully thought of as a term-introducing theory which implicitly defines terms such as ‘believe’, ‘want’ and ‘desire’ (e.g., Lewis 1972). Braddon-Mitchell and Jackson (2007), for example, follow this line and argue that the existence of folk psychological rules ‘does not, of course, mean that we must have a theory […] explicitly worked out in our minds, but somehow hidden from view and guiding our actions from its hiding place. Rather, it means that our responses to situations and our [folk psychological] judgments […] are governed in most cases by our existing networks of interrelated powers of discrimination' (p.63).

Of course, the question is what such an account of the 'existing networks of interrelated powers of discrimination' looks like - this is what an explanation of our folk psychological capacities should amount to. But Braddon-Mitchell and Jackson do not touch this question; they only argue that folk psychological rules can, in principle, be distilled from our common-sense use of psychological vocabulary. Hutto (2008a) rightly objects that we should not confuse this with the idea that those rules could *explain* the structural basis of folk psychology or that they are responsible for its genesis. In fact, most proponents of externalist TT are silent about issues of acquisition. Some of them have argued that instead of a futile search for the internal mechanisms of a theory of mind, we need to investigate our 'naïve' experience of social interaction: 'the psychological theory through which the concept of belief is introduced is a deeply tacit one. We must therefore look to common assumptions about belief reflected in our naïve use of belief to achieve any measure of success in the theory's articulation' (Zimmerman 2007, p.63).

What is interesting about these proposals is the attempt to vindicate the existence of folk psychological principles by appealing to the *social practice* in which they are articulated. In fact, I very much agree with proponents of externalist TT insofar they argue for an account of intersubjectivity that goes *beyond* the individual mind. However, although I applaud the suggestion to take a closer look at our everyday intersubjective engagements, I don't think this reveals how 'the theoretical principles do their work' and 'guide our mindreading activities'. On the contrary, I believe it provides us with a very different story about intersubjectivity (cf. chapter 5). However, even if it *would* lead to the uncovering of a deeply tacit theory, this by itself is certainly not sufficient to comfort those who are still worried about its context-sensitive application.

Moreover, the appeal to social practice can also be used to mount an extra argument *against* both externalist and internalist versions of TT. For example, in their evaluation of the myth of Jones and its significance for TT, Stich and Ravenscroft (1996) point out that, as Sellars tells the story, Jones self-consciously develops a folk psychological theory and explicitly teaches it to his compatriots. But Stich and Ravenscroft observe that nothing like that seems to go on in our current social practice: 'We don't explicitly teach our children a theory that enables them to apply mental terms to other people. Indeed, unlike Jones and his friends, we are not even able to state the theory, let alone teach it. If you ask your neighbor to set out the principles of the theory of the mind that she has taught her children, she won't have the foggiest idea what you're talking about' (pp.121-2). A similar argument against TT is made by Goldman (1989), who also wonders how children might get a grip on a theory as complex and sophisticated as the one that TT attributes to them: 'One possible mode of acquisition is cultural transmission (e.g. being taught them explicitly by their elders). This is clearly out of the question, though, since only philosophers have ever tried to articulate the laws, and most children have no exposure to philosophers' (pp.167-8). This brings us to a broader, more encompassing phenomenological argument against TT.

1.3 Where is the theory in TT?

The argument from phenomenology

Shaun Gallagher (2001, 2004) has argued that if the kind of theory-driven mindreading promoted by TT is central to social practice, then we should at least have some awareness of the fact that we are applying folk psychological rules when we try to read the mental states of others. However, there does not seem to be any phenomenological evidence for this, that is, there is no experiential evidence that we use theoretical principles when we are interacting with other persons. According to Gallagher, TT explanations of intersubjectivity in terms of mindreading presuppose that our encounters with others crucially depend on the ability to take a *third-person theoretical stance* in order to explain and predict their behavior. But taking such a theoretical stance, he argues, is a very specialized and relatively rare mode of social interaction, characterized by its reliance on an observational attitude and a lack of actual interaction. If we look at the 'phenomenological evidence' and pay attention to our daily life experience 'it seems likely that this explicit kind of meta-cognitive theorizing, although possible for the adult human, is not our everyday practice; it is not the way we think of ourselves or of others' (Gallagher 2004, p.202). This is what he calls 'the simple phenomenological argument'. Gallagher acknowledges that sometimes we *do* take a theoretical stance towards others, for example, in speculative discussions about third persons, or in situations when our interactions with others break down and we have trouble understanding them. However, these cases are the exception rather than the rule. Normally, intersubjectivity does not involve a 'detached or abstract observational stance', since our understanding of others 'is poorly described as involving the formulation of a theoretical hypothesis' (ibid.).

A similar critique against TT has been launched by Matthew Ratcliffe (2006), who argues that social interaction is 'seldom, if ever, a matter of two people assigning intentional states to each other […] Self and other form a coupled system rather than two wholly separate entities equipped with an internalized capacity to assign mental states to the other. This applies even in those instances where one might seem to adopt a "detached" perspective towards others' (p.31). Ratcliffe argues that folk psychology is an artificial creation of certain philosophers who have failed to attend closely enough to our real social practices, which operate in quite different ways. 'All I claim is that over the last fifty years, certain philosophers of mind and cognitive scientists have got into a bit of a muddle about intersubjectivity, and that the description of interpersonal understanding which they tend to adopt should be rejected' (2007, p.23). According to Ratcliffe, folk psychology is 'a misguided reification of abstractions that has no place in social reality' (ibid.).[8]

Although the strength of these phenomenological arguments lies in their straightforward appeal to our 'normal' experience of intersubjectivity, this is also their weakness (see chapter 2.1). Claims about what counts as an accurate phenomenological description of everyday social interaction are hotly disputed and difficult to resolve. At the same time, however, this by itself is already sufficient to block an explicit TT approach to intersubjectivity.

The appeal to tacit theory

Theory theorists usually try to parry these phenomenological arguments by going 'underground', arguing that the folk psychological rules they have in mind are drawn upon

[8] Notice that there are actually two different phenomenological arguments at play here: one against the TT interpretation of mindreading (Gallagher's), and one against mindreading more in general (Ratcliffe's). It is somewhat confusing that Ratcliffe uses the more restrictive term folk psychology instead of mindreading, given that besides TT, he aims to criticize other accounts of mindreading as well.

tacitly. Gopnik (2003) for example, suggests that 'the kinds of theory formation we see in children, the kind that lead to everyday knowledge do not, on the face of it, seem to be consciously accessible [...] In particular, children may not consciously assess evidence and consider its impact on theories' (p.247). And Crane (2003) also suggests that the theoretical rules or routines postulated by TT 'need not be explicitly known by us – that is, we need not be able to bring this knowledge to our conscious minds. But this unconscious knowledge, like the mathematical knowledge of Meno's slave [...] is none the less there. And it explains how we understand each other, just as (say) unconscious or "tacit" knowledge of the linguistic rules of grammar explains how we understand language' (p.67).[9]

If we employ the folk psychological principles necessary for mindreading in a *tacit* way, then what we experience or seemingly experience during social interaction is arguably not a good guide for what is 'really' happening in such cases. Because phenomenology is in principle unable to determine what is going on at the *unconscious* level, it cannot rule out tacit theory. However, in making this move, TT implicitly seems to concede the point that theory-driven mindreading fails as an adequate characterization of our everyday social exchanges. But things are more complicated than this. What TT typically concedes is that the phenomenological objections are correct only insofar as our *experience* of intersubjectivity is concerned: we are normally not conscious of attributing theoretically structured belief-desire pairs. But when the question is *what it is that we do* in order to make sense of others, the TT answer is still very much framed in theoretical terms: in some way or other, we attribute belief-desire pairs to them for the purpose of behavior explanation or prediction - if not consciously, then subconsciously. Thus, as Gallagher (2004) points out, advocates of tacit TT are still committed to claims about what happens at the personal level of social interaction. Hutto (2004) confirms this, observing that what is still implicitly assumed is that 'the main business of commonsense psychology is that of providing generally reliable predictions and explanations of the actions of others. In line with this, it is also generally assumed that we are normally at theoretical remove from others such that we are always ascribing causally efficacious mental states to them for the

[9] The reference to Plato is interesting here, because it is possible to interpret the Meno not only as the first formulation of the problem of knowledge, but also as the first (broadly) rationalist solution to it in terms of innateness. So is the analogy with the tacit rules of grammar, which shows TT's debt to Noam Chomsky.

purpose of prediction, explanation and control' (p.548).[10] The fact that this assumption about the nature of folk psychology is subsequently fleshed out in terms of tacit mindreading routines reveals that there is an important assumption of *isomorphism* at play here: an isomorphism between the sub-personal level of explanation and the personal level of description. But this assumption is questionable (cf. Gallagher 1997, Millikan 1993).[11] In particular, it turns out to be notoriously difficult to spell out the particular contents of tacit beliefs and/or desires. This has lead to a serious discussion about the very idea of locating (non)propositional content and attitudes at the sub-personal level (cf. Menary 2006, Hutto 2008). More in general, the question is whether it makes sense to apply concepts at sub-personal levels that were originally coined at the personal level.

Despite these obvious and legitimate worries, proponents of internalist TT maintain that the idea of tacit theorizing can and should be cashed out in terms of the cognitive neuropsychological processes of individual agents. They argue that instead of trusting our unreliable everyday experience, we should pay attention to certain scientific experiments that support their TT account of intersubjectivity. This is an interesting suggestion, and a closer look at the empirical evidence for tacit TT is certainly part of this chapter's program. But let us first consider an alternative way to make sense of the tacit folk psychological rules that are supposed to guide our social engagements.

Some theory theorists have argued that we need to postulate an *intermediate* level of intersubjective processing, an additional level of discourse between the phenomenological and the physiological that describes the way mindreading processes are guided by folk psychological principles from a *functional* perspective. Stich (1983), for example, has made a case for a *syntactic theory of mind* (STM). The core idea behind his proposal (what

[10] See, for example, Bogdan (1997, p.105), Botterill (1996, p.107) and Carruthers (1996, p.24).

[11] Contemporary neuroscience increasingly demonstrates that assumptions of isomorphism between the personal and the sub-personal level are seriously mistaken. Take the assumption of *spatial* isomorphism. The fact that a subject experiences a brighter patch as to the left of a darker patch, for example, does certainly not justify the conclusion that the neural activity responsible for the greater brightness of this left patch therefore also must occur to the left of the activity responsible for that of the dark patch. Or consider *temporal* isomorphism. Dennett (1991) has pointed out that Libet's work on backward referral in time suggests that there might very well be no isomorphism between the temporal structure on the neurobiological level and the serial structure of that which is represented on the conscious level. Gallagher (1997) stresses this point as well, and also makes an additional argument against *quantitative* isomorphism, referring to the well-known fact that the brain processes a larger quantity of information about environmental features than we become conscious of in perception (see also Marcel 1983).

makes it syntactic) is that folk psychological knowledge cannot be mapped directly onto our individual brains, as modular theory theorists want to have it, but first needs to be specified in terms of its formal or syntactic structure. This syntactic structure subserves the beliefs and desires we employ in our daily social interactions, but it does not address their specific folk psychological *contents*. 'Cognitive theories which cleave to the STM pattern treat mental states as relations to purely syntactic mental sentence tokens, and they detail the interactions among mental states in terms of the formal or syntactic properties of these tokens' (p.9). Stich thinks that too much attention to the contents of mental states imposes damaging restrictions on the scope and methods of cognitive psychology. Cognitive psychology seeks causal explanations of behavior and cognition, and the causal powers of mental states are determined by their *syntactic* properties.

A recent product of this line of thinking is the Early Mindreading System (Nichols and Stich 2003). The Early Mindreading System is embedded in a larger Basic Cognitive Architecture, and consists of a trio of mechanisms (fig. 1.1).

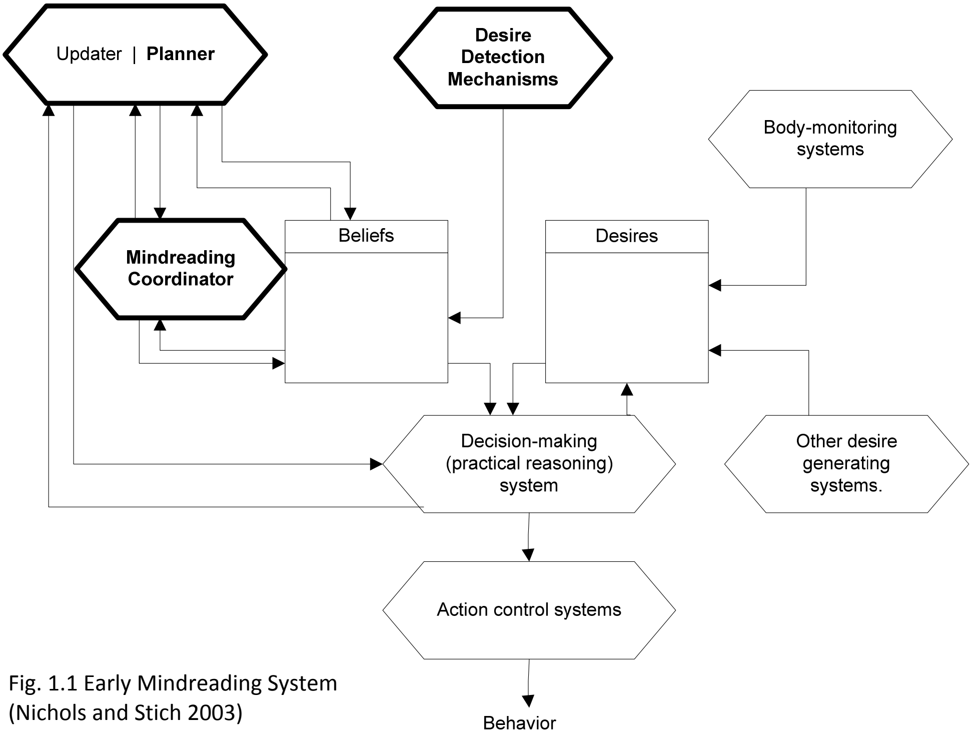


Fig. 1.1 Early Mindreading System
(Nichols and Stich 2003)

The first mechanism is actually a cluster of mechanisms, labeled ‘desire detection mechanisms’ (p.78). These mechanisms infer the desires of other people and feed them into a second mechanism: the ‘planner’. This mechanism plays an essential role in the generation of actions, and its only function is to calculate which actions lead to the satisfaction of these particular desires (whether the mindreader himself has the desire in question or not). In the Early Mindreading System this process is still somewhat dysfunctional, since the planner is not yet able to take into account all the relevant information about others. What is also missing at this stage is information about the beliefs of others. The planner mechanism simply assumes that the other has the same beliefs as the mindreader (p.80). The third mechanism is the Mindreading Coordinator. One important function of the Mindreading Coordinator is to turn on the desire detection mechanisms when additional information about others’ desires is necessary. Once this information is acquired, the Mindreading Coordinator sends the mindreader’s beliefs about the desires of the other to the planner mechanism (p. 81). In the final step, it turns the output of the planner mechanism into a belief about the other’s intentions or goals. The Mindreading Coordinator also takes care of a number of miscellaneous tasks, such as ‘cleaning up’ the old beliefs when beliefs about the other’s desires have changed.

I already remarked that TT has problems when it comes to explaining how we are able to specify the conceptual *contents* of the beliefs and desires of the people we try to understand. This is required if we want to apply our mindreading skills in a context-sensitive manner. The Early Mindreading System seems to be able to circumvent this problem, because it facilitates a very basic kind of mindreading that requires only the slightest of conceptual understanding. Although Nichols and Stich assume that early mindreaders (that is, very young, non-verbal children) already have beliefs, they are thought not yet to have mastered the *concept* of belief. It is the special kinds of beliefs that they have and what they do with them that yields a practical understanding of the intentions and goals of others. However, as Hutto (2007a) remarks, it is unclear what ‘having a belief’ comes to at the sub-personal level. Moreover, it is unclear how the Early Mindreading System could work at all if such a belief would have no content *whatsoever*. What is needed is a kind of tacit belief that comes with non-linguistic representational content, but Hutto convincingly argues that such a notion is unintelligible. I will provide a more detailed elaboration of this claim later on in this chapter.

1.4 Scientific evidence for TT

Theorizing chimps

Many theory theorists argue that the major tenets of their position are based on well-designed scientific experiments. An important landmark in the experimental history of TT is the publication of Premack and Woodruff's (1978) article 'Does the chimpanzee have a theory of mind?' The article starts with a general declaration of commitment to TT. Premack and Woodruff claim that each human being has a theory of mind, which means that he 'imputes mental states to himself and to others [...] A system of inferences of this kind is properly viewed as a theory, first, because such states are not directly observable, and second, because the system can be used to make predictions, specifically about the behavior of other organisms' (p.515). Then follows the interesting question: is it possible that chimpanzees possess a theory of mind that is not markedly different from our own? Premack and Woodruff report an experiment in which they showed chimpanzees videotapes of humans in problem situations that the animals could presumably understand (e.g., trying to retrieve bananas that are placed above their reach). The animals were then shown a series of photographs, one of which depicted a possible solution to the problem (e.g., a moveable box that allowed the human to reach the bananas). According to Premack and Woodruff, the fact that the chimpanzees tended to choose the best answer meant that they were able to adopt the perspective of the person in the video. And this, they argue, implies that chimpanzees have a theory of mind. Premack and Woodruff also suggest that it might be interesting to study theory of mind in other populations: 'Although here we have talked only about the chimpanzee [...] are at least some retarded children deficient in specifically this form of theory building? What is the developmental course of such theory building in the normal child?' (pp.525-6).

Premack and Woodruff's suggestion has been very influential, and their article set the stage for a major episode in research on theory of mind in children. As Gopnik (1993) observes, 'in the last few years there has been an explosion of interest in children's ideas about the mind' (p.3). In a similar fashion (but on a more critical note), Reddy and Morris (2004) remark that it 'is difficult to write today about understanding people without reference to the words "theory of mind". An incredible 1 percent of academic publications in psychology in 2003-4 that refer to infants or children also refer to the term "theory of mind". And the manner in which the term is used is awesomely matter-of-fact-with a taken-for-grantedness hitherto reserved for those other staples of psychology such as "growth spurt", "toilet training", "short-term memory" and "secure attachment"' (p.647).

The false belief test

The peer commentary that followed Premack and Woodruff's article showed that simply predicting the action of others, as the chimpanzees were asked to do, was not sufficient to distinguish between 'mindreaders' and 'behaviorreaders'. Dennett (1978) in particular laid out some of the difficulties in making this distinction and offered some empirically-friendly suggestions aimed at teasing them apart. According to him, a key component absent from Premack and Woodruff's experiments was not only a measure of *false* belief attribution, but also a measure of false belief attribution in a *novel* situation. The former is required to rule out the possibility that subjects simply choose on the basis of their own beliefs instead of the beliefs of others, and get it right by accident. The latter is required to rule out a behaviorist explanation in terms of experienced regularities. Dennett suggested a scenario suitable for young children, in which Punch had a mistaken belief about the location of Judy. Wimmer and Perner (1983) modified this scenario slightly, and voilà: a cottage industry of experiments with young children was born.

The core idea behind the false belief test is that children need to demonstrate the ability to recognize that others may have *false* beliefs plus the ability to *predict* their behavior on the basis of these beliefs. There are more or less difficult variations of the false belief test. A very popular one is the 'Sally-Anne' test, which goes as follows. First, the child is shown the scenario illustrated below (fig. 1.2), which can be enacted by puppets or real people. Then, the child is asked where Sally will look for her ball. To answer this question correctly, the child must realize that Sally has not seen the ball being moved and, therefore, that Sally *falsely believes* that the ball is still in the basket.

Results show that 3-year-olds fail this task because they do not understand that Sally has a false belief about the location of the object. Four-year-olds, by contrast, typically answer correctly and are thus capable of distinguishing between 'how things really are in the world and what other people may falsely believe about such things' (Gallagher 2004, p.199).

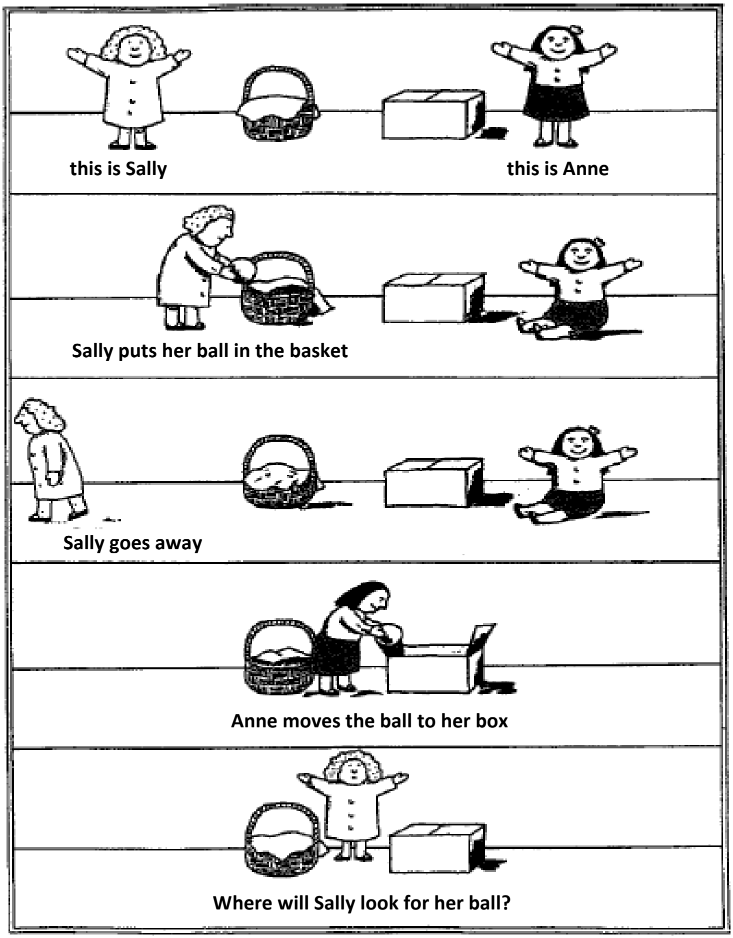

this is Sally
this is Anne
Sally puts her ball in the basket
Sally goes away
Anne moves the ball to her box
Where will Sally look for her ball?

Fig. 1.2 The Sally-Anne False belief test

Another example is the ‘Smarties’ test. Children are presented with a candy box, which is actually full of pencils, and then they are asked what they think other people will think is in the box. Three-year-olds consistently say that other people will think there are pencils in the box, and they continue to make this error when they see them responding to the box with surprise - even when they are explicitly told about their false beliefs (Perner et al. 1987, Moses and Flavell 1990, Wellman 1990).

False belief tests similar to the ones described above have also been used to uncover the neurobiological processes underlying our mindreading abilities.[12] In a number of experiments, evidence was found for a neural network comprising the medial prefrontal cortex, the superior temporal sulcus (especially around the temporo-parietal junction) and the temporal poles adjacent to the amygdala (cf. Fletcher et al. 1995, Saxe and Kanwisher 2003, Vogeley et al. 2001). Other neuroimaging studies have also implicated the frontal cortex in this network (cf. Happe et al. 2001, Rowe et al. 2001, Stone et al. 2001, Gregory et al. 2002).

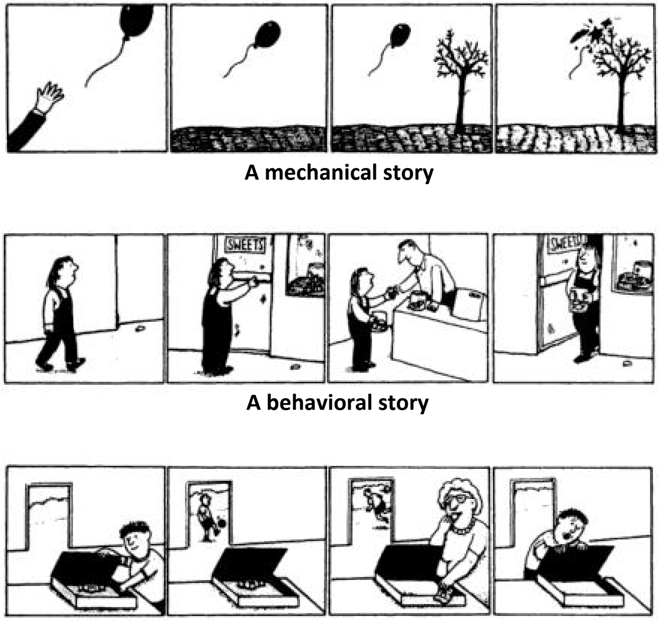
According to proponents of TT, the results on false belief tests show that children typically appear to cross a theory of mind threshold between the age of 4 and 5. Before this age, they are not yet able to understand that the beliefs of another person may be false. But between the age of 4-5, children develop the basics of a theory of mind that enables them to attribute 'first-order beliefs' to others that are different from their own beliefs. This theory of mind develops and gets increasingly sophisticated as children mature. Between the age of 6 and 7, children acquire the ability for 'second-order belief attribution' and become able to 'think about another person's thoughts about a third person's thoughts about an objective event' (Baron-Cohen 1989, p.288).

In cases of autism, however, false belief tests show that children have trouble in acquiring the ability for first and second-order belief attribution. This was first noticed by Baron-Cohen et al. (1985) in the article 'Does the autistic child have a Theory of Mind?' The investigators reported an experiment in which the 'Sally-Anne' false-belief task was administered to a group of autistic children, a group of children with Down syndrome, and a group of normal pre-school children. All these children had a mental age of above 4 years. The experiment showed that 80 percent of the autistic children failed the false belief task. By contrast, 86 percent of the Down syndrome children and 85 percent of the normal preschool children passed the test. On the basis of these percentages, the experimenters concluded that autistic children have serious difficulty recognizing the significance of false belief.

In another experiment, Baron-Cohen et al. (1986) gave the subjects scrambled pictures from comic strips with the picture already in place. The subjects were supposed to

[12] There are also neuroimaging studies that have investigated the attribution of other mental states than beliefs, such as desires and goals (Decety et al. 2002, Chaminade et al. 2002, Saxe et al. 2004).

put the strips in order to make a coherent story and also tell the story in their own words. There were three types of stories: mechanical, behavioral, and mentalistic stories (fig. 1.3).



A mechanical story

A behavioral story

A mentalistic story

Fig. 1.3 Three types of picture sequences

All the autistic children ordered the pictures in the mechanical script correctly and used the right kind of language when telling the story; for instance, ‘the balloon burst because it was pricked by the branch’. They also dealt adequately with the behavioral script, which could be told without reference to mental states. But the vast majority of them could not understand the mentalistic stories. They put the pictures in jumbled order and told their stories without any attribution of mental states. These and other findings led Leslie and Frith (1988) to suggest that autistic children might be specifically impaired in their capacity for meta-representation, which in turn impedes the development of a theory of mind.

Neuroscientists have tried to trace the neurobiological roots of this impediment. Castelli et al. (2002), for example, PET-scanned autistic and normal subjects while they were watching animated sequences. The animations depicted two triangles moving about on a screen in three different conditions: moving randomly, moving in a goal-directed fashion (chasing, fighting), and moving interactively with implied intentions (coaxing, tricking). The last condition frequently elicited descriptions in terms of the mental states that viewers attributed to the triangles. The autistic subjects gave fewer and less accurate descriptions of these animations, but equally accurate descriptions of the other animations. While viewing animations that elicited mindreading, in contrast to randomly moving shapes, the normal subjects showed increased activation in the neural network described above (the medial prefrontal cortex, the superior temporal sulcus at the temporo-parietal junction and temporal poles). The autistic subjects showed less activation than the normal subjects in all these regions. However, one additional region, the extrastriate cortex (which was highly active when watching animations that elicited mindreading) showed the same amount of increased activation in both groups. In the group with autistic subjects, this extrastriate region showed reduced functional connectivity with the superior temporal sulcus at the temporo-parietal junction, an area associated with the processing of biological motion as well as with mindreading. The experimenters concluded that this indicated a physiological cause for the mentalizing dysfunction in autism, namely, a bottleneck in the interaction between higher-order and lower-order perceptual processes.

A question of interpretation

The crucial question is what the above findings tell us about children's ability to understand others. Do they support a TT explanation of intersubjectivity in terms of mindreading? Bloom and German (2000) have warned us that we should be very careful in interpreting the findings resulting from the false belief test, since it is an 'ingenious, but very difficult task that taps (only) one aspect of people's understanding of the minds of others' (p.30). This point is also made by Gallagher (2004), who argues that false belief tests are designed to capture very specialized cognitive abilities that allow us to predict and explain the behavior of others in a third-person context. But these abilities 'put us in an observational mode and do not capture the fuller picture of how we understand other people' (p.204). Gallagher (2005) claims that there are at least three factors that limit the conclusions that can be drawn from false belief tests in order to support TT:

1) The experiments explicitly test for the *specialized* cognitive activities of explaining and predicting.
2) The experiments involve *third-person observations* rather than *second-person interactions*.
3) The experiments involve *conscious* processes and do not address theory-of-mind mechanisms that operate *non-consciously*.

Since proponents of TT assume that intersubjectivity is primarily about the prediction and explanation of behavior in a third-person context (and thus are committed to 1 and 2), the question arises whether their appeal to false belief tests is not rather a *self-fulfilling prophecy*. Stich and Nichols (1992), for example, suggest that 'the explanation of the data offered by the experimenters is one that presupposes the correctness of the theory-theory' (p.62). And Ratcliffe (2007) also points out that 'the very design of the task and the importance ascribed to it simply presupposes that a detached ability to assign intentional states is central to interpersonal understanding' (p.228). In other words, it is by no means clear that the specialized cognitive abilities that are captured by the false belief test are fundamental to action understanding.

Another problem is that TT generally assumes that the ability to understand false beliefs is acquired *across the globe*, i.e. universally. However, several cross-cultural experiments with children from non-Western cultures indicate that these children fail to perform on standard false-belief tests as readily or with the same proficiency as Western children do (Vinden 1996, 1999, 2002; Lillard 1997, 1998; Garfield et al. 2001). The studies by Vinden, for example, reveal significant differences in the understanding of belief between children of certain cultures. 'The response patterns vary from culture to culture, with the Western children the only ones who were at ceiling on all questions' (1999, p.32).

What is also very problematic about the false belief test is that, because of its narrow focus on third- person contexts of action understanding, it strips away structures of interaction that are constitutive of our everyday second-person encounters. Bloom and German (2000) remark that 3-year-olds often pass more 'pragmatically natural' variants of the false-belief test with simpler or more specific questions. They suggest that younger children do not have the blanket ignorance of alternative perspectives on the world that failure on the false belief test may suggest. This is supported by naturalistic home-based family studies, which show that children are usually well-attuned to other people's states of ignorance, their emotions and desires, and demonstrate sufficient understanding that other people's likes and desires may be different from the child's own (cf. Dunn 1988).

Gallagher (2005) has argued that there are false belief test set-ups that might address the lack of second-person interaction to a certain extent. An experiment by Wimmer et al. (1988), for example, had two children face each other while they were answering questions about what they knew, or about what the other child knew about the contents of a box into which one of them had looked. This seems to come a lot closer to second-person interaction. What the experiment showed is that children of 3 and 4 year answer correctly about their own knowledge, but incorrectly about the other child's knowledge, even when they know that the other child has looked into the box. However, Gallagher points out that even here the children are still not really *interacting:* 'the questions are posed by the experimenter (with whom the children are interacting) but they call for third-person explanation or prediction of the other person with whom they are not interacting' (2005, p.219).

The above experiment is interesting because it shows that there might be a difference between children's knowledge of the *other* mind and the knowledge they have of their *own* mind. According to many theory theorists, these kinds of knowledge can only differ in *degree* because they are derived from the same folk psychological theory. In terms of development, this means that children would acquire self-knowledge around the same time they acquire knowledge of others, and encounter the same difficulties in both cases.

Gopnik and Meltzoff (1994) argue that most of the developmental evidence indeed points in this direction. They claim that the evidence suggests that there is an extensive parallelism between children's understanding of their own mental states and their understanding of the mental states of others: 'In each of our studies, children's reports of their own immediately past psychological states are consistent with their accounts of the psychological states of others. When they can report and understand the psychological states of others, in the cases of pretense, perception and imagination, they report having had those psychological states themselves. When they cannot report and understand the psychological states of others, in the case of beliefs and source, they do not report that they had those states themselves' (pp.179-80). I already mentioned the 'Smarties' test, in which children are presented with a candy box full of pencils. Gopnik and Astington (1988) have shown that 3-year-old children not only predict that others will think there are pencils in this candy box (without having looked into it), but also that they make the same error when they are asked about their *own* immediately past false beliefs. In this case, they report that they *already thought* that there were pencils in the box. According to Gopnik (1993), this proves that there is no such thing as a privileged access to our own mind.[13]

Although these findings are certainly of importance, the question is how we should *interpret* them. Some theory theorists have suggested that a possible difference between self and other knowledge ultimately comes down to a difference in *mental processing.* Children use an 'answer check procedure' in order to answer questions about their own knowledge. According to such an account, children need to check whether they themselves know what is in the candy box and this involves something like a meta-representational introspection (cf. Leslie 1988). However, Gallagher (2005) has argued that there is a much more likely and parsimonious explanation of what happens in these cases: 'their answer about what they know is based simply on looking inside the box rather than looking inside their own mind. The child looks inside the box and is then asked whether she knows what is in the box. Her positive answer is based on the fact that she just saw what was inside the box, rather than on an introspective discovery of a belief about the contents of the box' (pp.219-20). Moreover, even if we would grant the importance of mentalistic procedures in the very specific context of the false belief test, the question still remains whether there is any phenomenological evidence for the claim that we consciously employ these procedures in *other* contexts as well. This brings me to the third limitation mentioned by Gallagher.

Sometimes, theory theorists admit that false belief tests only capture a small part of how we understand others. And sometimes, they are also willing to accept that false belief tests are of no use in supporting TT interpretations of 'low-level' implicit forms of intersubjectivity, since they address the kind of social understanding of which children are *conscious*. However, these theory theorists still maintain that low-level forms of intersubjectivity are thoroughly theoretical, because there are many *precursors* to the explicit attribution of false belief that is measured by the false belief test. Of course,

[13] But see also the passage in Gopnik (1993) where she argues that 'One possible source of evidence for the child's theory may be first-person experiences that may themselves be the consequence of genuine psychological perceptions. For example, we may well be equipped to detect certain kinds of internal cognitive activity in a vague and unspecified way, what we might call 'the Cartesian buzz' [...] Our genuinely special and direct access to certain kinds of first-person evidence might account for the fact that we can draw some conclusions about our own psychological states when we are perfectly still and silent' (p.11). This clearly clashes with her earlier remarks.

different experiments are needed to test for this kind of implicit false belief attribution. A popular reference in this respect is the violation-of-expectation experiment by Onishi and Baillargeon (2005), which attempted to show that infants at the age of 15 months already have a rudimentary understanding of the false beliefs of other people. In this experiment, infants were first familiarized with an adult hiding a toy in one of two locations, and then presented with scenes where the toy was moved without the adult's knowledge. Subsequently, they were shown scenes of the adult searching for the hidden toy either where she falsely believed it to be, or where it was actually located. Onishi and Baillargeon found that infants reliably looked longer at what they called the 'unexpected event', where adults searched at the correct location despite their false belief about where the toy was hidden. This means, according to them, that the infant in fact expected the adult to search for the toy where she *believed* it had to be located (cf. Clements and Perner 1994). Should we interpret these findings as providing evidence for an implicit precursor to an explicit folk psychological theory?

Belief-desire psychology in low-level action understanding?

Throughout this book we will encounter many scientific experiments that can be interpreted in such a way as to support TT, while in fact a far more parsimonious explanation is available. One of the main aims of this book is precisely to provide such an explanation. At the same time, however, the simple fact that it is possible to come up with *different* explanations should urge us to treat the evidence resulting from these kinds of experiments with extreme caution. This casts doubt on the idea that the evidence by itself is sufficient to decide the debate in favor of one position or the other. In this respect, I fully agree with Stueber (2006) who suggests that ‘empirical considerations about underlying mechanisms alone - especially neurobiological mechanisms - as important as they are for understanding of folk psychological abilities, can never decide the issue’ (p.100).

Explicit versions of the false belief test clearly show that something new and important happens at the age of 4 years, and that this something is somewhat consistent with certain assumptions of TT. However, since these tests are designed to capture a very specialized mode of intersubjectivity, they cannot be used to validate a TT approach to intersubjectivity *in general.* This is not because the evidence is lacking, but it rather has to do with the proper interpretation of this evidence and the questioning of a certain picture of intersubjectivity that is presupposed by TT. In implicit false belief tests of the Onishi and Baillargeon kind, it also concerns the appropriate *level of explanation*. Suppose we grant TT that our understanding of others is facilitated by a *tacit* folk psychological theory. The question still remains whether it is possible to map belief-desire processes directly onto the sub-personal level, using personal level vocabulary as if nothing has changed.[14] At the very least, we should carefully explain what we mean by 'tacit' beliefs and/or desires.

We only have to consider what theory theorists *themselves* say about the notion of belief at the personal level of social understanding to find out that this is importantly different. Here, the focus is primarily on belief as a cognitive attitude that aims at truly representing how things stand with the world. A belief is *about* certain states of affairs in the world (and thus intentional), and has the virtue that it can be *verified*. Understanding a false belief implies that one can distinguish between a true and a false descriptions of a state of affairs in the world, and also that one has the ability to demonstrate how a belief about this state of affairs can be false. And this, in turn, presupposes that one has learned the correct *procedures* to do so (this is exactly what is measured by the explicit false belief test). What Onishi and Baillargeon's violation-of-expectation experiment shows is that children are intentionally directed at aspects of their environment, and that *we* (or better: the experimenters) are able to describe this in terms of truth-evaluable beliefs that are part of a larger theoretical framework. But that does certainly not mean that these children themselves have full-blown beliefs or use a theory to coordinate their behavior, any more than planets use Newtonian laws in order to conduct their business (to borrow an example from Hutto).[15]

[14] Dennett (1969) states the dilemma clearly: 'When we have said that a person has a sensation of pain, locates it, and is prompted to act in a certain way, we have said all there is to say within the scope of this (personal-level) vocabulary. We can demand further explanation of how a person happens to withdraw his hand from the hot stove [...] but if we do this we must abandon the explanatory level of people and their sensations and activities and turn to the sub-personal level of brains and events in the nervous system. But when we abandon the personal level in a very real sense we abandon the subject matter of pains as well [...] for our alternative analysis cannot be an analysis of pain at all, but rather of something else - the motion of human bodies or the organization of the nervous system' (pp.93-4).

[15] If the only requirement for folk psychological competence is that one's behavior can be described in a structural, truth-evaluable way, then almost anything deserves this title: not only human beings, but also animals, tornados and thermostats. Theory theorists seem to be confronted with the following dilemma here: either they should grant folk psychological competence to anything which can respond discriminatively to classes of objects, or else they should explain what more is needed for folk psychological competence.

It pays to consider Hutto's (2007) distinction between *intentional* attitudes and *propositional* attitudes here. Hutto argues that nonverbal animals and preverbal infants display intentional attitudes insofar as they intentionally respond to certain aspects of their environment. He coins the term 'biosemiotics' to characterize the kind of non-verbal thinking he has in mind, which basically boils down to Millikan's biosemantics without representationalism (Millikan 1993, 2004). Although intentional attitudes do not involve nor implicate truth-conditional *content*, they can still account for an impressive range of sophisticated, non-linguistic activities. Hutto claims that intentional attitudes already provide children with the necessary means to interact with the world in a meaningful way, long before they develop a basic understanding of propositional attitudes. The latter are exclusively employed by those beings that have mastered certain linguistic constructions and practices, including the ability to represent and reason about complex states of affairs in truth-evaluable ways. Following this line of thought, we might argue that Onishi and Baillargeon have shown that infants at the age of 15 months display *intentional attitudes*, but this certainly does not prove they already master *propositional attitudes*.

Even if we do not want to buy into this distinction, it could still be remarked that Sally-Anne false belief tests and violation-of-expectation experiments only satisfy those theory theorists who employ a 'vegetarian' concept of folk psychology. At best, these findings explain the acquisition of the understanding of a propositional attitude *in isolation*. But why would theory theorists think that such an isolated 'understanding' of false belief is enough for folk psychology? It is remarkable that many proponents of TT, *who have explicitly committed themselves to belief-desire psychology*, are not in the least troubled by the fact that the evidence they appeal to only supports a (by their own standards) completely oversimplified picture of folk psychology. Hutto (2007a) argues that in order to practice folk psychology in a meaningful way, children need more than an isolated understanding of the propositional attitudes *an sich:* 'Knowing that children manage to pass false-belief tests, reliably enough, at a certain age under very particular experimental conditions, gives no insight into the extent of their understanding of that concept in other contexts' (p.26). This is because in order to make sense of an action as performed for a *reason*, 'it is not enough to imagine it as being sponsored by a singular kind of propositional attitude; one must also be able to ascribe other kinds of attitudes that act as relevant and necessary partners in motivational crime' (ibid.). Folk psychology *stricto sensu,* as Hutto labels it, at the very least involves the ability to make sense of another person's actions using belief-desire propositional attitude psychology. Knowledge of how these propositional attitudes interrelate with one another ‘comprises what we might think of as the “core principles” of intentional psychology’ (p.29).

But Hutto argues that there is also another important requirement that needs to be met. Children also need to become familiar with the norm-governed possibilities for wielding folk psychology *in practice*, so that they can apply it sensitively – adjusting for relevant differences in particular cases by making allowances for a range of variables such as the person's character, circumstances, etc. As we have shown in the previous sections, this is a serious problem for TT, and it is therefore not surprising that its proponents usually don't bother to explain how children become able to do this. Hutto’s solution to the problem of context-sensitivity is to argue that folk psychology has a *narrative* as opposed to a *theoretical* structure. According to his ‘narrative practice hypothesis’, the main developmental route through which children become familiar with the background norms for wielding folk psychology in practice is by being exposed to ‘folk-psychological narratives’. The defining feature of these narratives is that they reveal how beliefs and desires (and other propositional attitudes) interrelate and conspire to form reasons for action. I discuss this proposal in greater detail in chapter 5.

1.5 A number of pressing TT problems

A short summary

So far we have encountered a number of problems that accompany TT explanations of intersubjectivity. These problems are ‘internal’ in the sense that they arise when one accepts a TT picture of intersubjectivity. If we assume that our meetings with other minds are facilitated by a folk psychological theory, then one of the first questions is: where does this theory come from? Despite their disagreement about the role of innateness in the acquisition of folk psychological content, all internalist versions of TT eventually appeal to innateness in their account of how we acquire the folk psychological rules that structure our mindreading activities. Externalist versions of TT, on the other hand, argue that these rules can, in principle, be distilled from our common-sense use of psychological vocabulary. However, attempts to articulate these putative laws or ‘platitudes’ have been notably weak, and externalist TT also lacks a developmental story.

All TT positions face serious difficulties in explaining how we acquire the background knowledge needed to sensitively apply our folk psychological theory in the large variety of practical contexts in which we supposed to exercise our mindreading skills. The appeal to innateness is a tempting solution to this problem, but we might wonder with Gopnik whether this actually amounts to an *explanation*.

Another problem concerns the *phenomenology* of intersubjectivity. In response to TT’s assumption that our social encounters are best characterized as theoretical predictions and explanations of behavior in third-person contexts, we might ask whether this fits the phenomenology of our everyday social life. The fact that such a lean picture of intersubjectivity is also presupposed by false belief tests severely limits the conclusions that can be drawn from their results. At best, TT might be able to explain a very specialized and relatively rare mode of social interaction.

There are also many conceptual problems. Insofar as proponents of TT appeal to tacit theory, the question is what it means to talk about folk psychological processes in terms of beliefs and desires at the sub-personal level. This requires not only a careful interpretation of the relevant empirical evidence that is brought forward to support this kind of tacit theorizing, but also a conceptual analysis of the notions of belief and desire that are claimed to be involved in these processes. The question is whether it makes sense to apply concepts at sub-personal levels that were originally coined at the personal level. But even if there is something to be said for the application of personal-level concepts at the sub-personal level, and even if some of the evidence *could* be understood in this way, its merits cannot be appreciated without regaining a clear understanding of what such sub-personal belief-desire processing is supposed to explain. In other words, the question still remains whether TT in fact provides us with a satisfying description of intersubjectivity at the personal level. As I have shown in this chapter, there are serious reasons to doubt this.

Together, these problems give us some reason to doubt the idea that intersubjectivity is best understood as the third-person explanation/prediction of behavior by means of a theory. For a more thorough critique of TT, however, we need to know more about its basic assumptions (chapter 3). And, of course, we might want to see what a healthy alternative would look like (chapter 4-5).

Folk psychology is false as a theory...

In this section I wish to address one more objection to TT. What is interesting about this objection is that it arises as soon as we affirm TT's basic assumption that folk psychology is indeed a *theory*. As noted earlier, it was Sellars who pointed out that the special epistemological status we attribute to certain claims is not based on privileged access, but on self-ascriptions that depend on an inherited and internalized theoretical framework. Sellars himself probably regarded this framework as empirically correct. However, if our knowledge of others is based on a folk psychological theory and in principle *falsifiable*, then this theory might be *false* as well.

This is the starting point for a radical critique of folk psychology initiated by Paul Churchland (1981, 1988). Churchland begins his argument by noticing that the mind/brain is a furiously active theorizer from the word go. He claims that 'the perceptual world is largely an unintelligible confusion to a newborn infant, but its mind-brain sets about immediately to formulate a conceptual framework with which to apprehend, to explain, and to anticipate that world [...] The furious conceptual revolution undergone by every child in its first two years is probably never equaled throughout the remainder of its life' (1988, p.80). He then goes to great lengths to demonstrate that this rapidly developing conceptual framework meets all the criteria of a theory, which eventually enables adult human beings to explain and predict the behavior and mental states of other persons. Churchland argues that our mature common sense psychological explanations can be construed as following a nomological-deductive pattern that is based on a web of interrelated, law-like generalizations of the following sort (cf. Sleutels 1994, p.48):

$(\forall x)(\forall p)(\forall q)$ {(x hopes that p) & (x believes that (if q then $\neg$p)) & normal circumstances $\rightarrow$ (x hopes that $\neg$q)}

$(\forall x)(\forall p)(\forall q)$ {(x believes that p) & (x believes that (if p then q)) & normal circumstances $\rightarrow$ (x believes that q)}

$(\forall x)(\forall p)(\forall q)$ {(x desires that p) & (x sees that $\neg$p)) & normal circumstances $\rightarrow$ (x is disappointed to find that $\neg$p)}

According to Churchland, these generalizations are fallible empirical hypotheses. The mental concepts they employ are defined by their place in the overall system of laws. They are the theoretical terms of a theoretical framework, and their meanings are fixed by the set of generalizations in which they figure. ‘Theoretical terms do not, in general, get their meanings from single, explicit definitions stating conditions necessary and sufficient for application. They are implicitly defined by the network of principles that embed them’ (1988, p.56). Theoretical terms primarily have a predictive and explanatory function, and Churchland argues that this is also their main value.

However, the above observations only serve to pave the way for a more important and provocative claim: that folk psychology is a radically *false* theory. Churchland argues that folk psychology offers us a ‘false and radically misleading conception of the causes of human behavior and the nature of cognitive activity’ (1988, p.43) and claims that 'the folk psychology of the Greeks is essentially the folk psychology we use today, and we are negligibly better at explaining human behavior in its terms than was Sophocles. That is a very long period of stagnation and infertility for any theory to display’ (1981, p.74). A future neuroscience is likely to have no need for notions such as beliefs and desires, and Churchland proposes to *eliminate* these concepts in order to make room for more precise and objective phenomena such as neurons and neural networks.[16]

...but folk psychology might not be a theory

The most effective way to counter Churchland’s eliminativist move is probably to agree with Stich (1983) that folk psychology is a ‘multi-purposes tool’ that is designed for various purposes, none of them scientific. Folk psychology does have practical value in real-life situations, but it gives no 'deep' explanation of our behavior. It stands to proper scientific psychology as cooking stands to chemistry. However, if this is true, i.e. if the application of scientific standards of theory evaluation to mindreading is misguided, then why should we think of mindreading in terms of *theory*? Perhaps mindreading has an entirely different

[16] Churchland (1989) gives a number of direct arguments against folk psychology that can be summarized as follows: (i) most folk theories have proved false, therefore it is unlikely that folk psychology will turn out to be true, (ii) folk psychology is an empirically and conceptually degenerating research program; as such, it deserves to be terminated, and (iii) there is a vastly superior competitor to folk psychology, namely, the new research program in cognitive neuroscience.

explanation. Perhaps it depends on *simulation*. This possibility will be explored in the next chapter.

Let me close by addressing what I think is a very *sensible* assumption which underlies the TT framework: that the meaning of folk psychological terms depends on their role in a larger network. This assumption goes against the view that these terms get their meaning by 'inner ostension' – by being directly associated with a specific quality of internal and privately experienced mental states. The latter idea is at the basis of the argument from analogy, according to which my knowledge of the other mind is indirect and analogical, an inference from my own case. Interestingly, however, it is also what fuels most theory of mind research. Wellman and Phillips (2001), for example, argue that children use the verb 'want' to 'refer to a person's internal state of wanting or longing to obtain an object, engage in action, or experience a state of affairs' (p.130). But this is certainly not in line with TT's claim that the meaning of notions such as 'belief' and 'desire' is fixed by their role in a larger conceptual framework. In other words, we should be careful in our evaluation of empirical research that is carried out under the heading 'theory of mind', since this is not necessarily compatible with the basic assumptions of TT.

Although we might agree with TT that the meanings of mental terms depend on their role in a larger network, it does not automatically follow that the network in question is a theoretical one and these terms thus have a theoretical status. As Hutto (2007, p.31) puts it: 'the mere fact that something has a framework structure does not entail that it is a theory [...] Ordinary games, such as cricket or chess, have rules, but these activities are not theoretically but conventionally grounded; they are well-established, regulated social practices. Folk psychology, too, has a frame-work structure, but it is neither a game nor a theory'.

2.

Simulation Theory

You know my methods in such cases, Watson. I put myself in the man's place, and, having first gauged his intelligence, I try to imagine how I should myself have proceeded under the same circumstances. In this case the matter was simplified by Brunton's intelligence being quite first-rate, so that it was unnecessary to make any allowance for the personal equation, as the astronomers have dubbed it.

- Doyle 1986

Folk psychology is simulation

Simulation theory (ST) has its starting point in the idea that everyday social interaction depends on the use of one's own mind as an internal model to understand the minds of others. Like Sherlock Holmes, our strategy to solve the mystery of the other mind involves putting ourselves in the other's shoes and imagining how we should ourselves have proceeded under the same circumstances. To understand the other person, we have to *simulate* the thoughts, feelings or behaviors that we would have in a similar situation.

The main objective of this chapter is to assess the strengths and weaknesses of ST as an approach to intersubjectivity. Obviously, such an assessment needs to be sensitive to the fact that there are various ways to further unpack the notion of 'simulation', resulting in different versions of ST, each with a different amount of philosophical baggage. Moreover, to do justice to these different versions of ST, we cannot avoid considering the complicated and traumatic relationship that ties them to their ancestor TT. What the early papers on ST (Gordon 1986, Heal 1986, Goldman 1989) had in common was a strong desire to move away from the over-intellectualized picture of social interaction offered by TT. ST was proposed as a solution to the problem of 'theory' in TT, and as such posed a direct challenge to the latter.[17] Theory theorists argued that our social engagements crucially involve mindreading, a procedure that allows us to explain and predict the behavior of our fellow human beings in terms of mental states such as beliefs and desires. But they also maintained that the success of this procedure depends on a folk psychological *theory* - a body of principles delineating how these beliefs and desires relate to perceptions, bodily expressions, (verbal) behavior and other mental states. Early proponents of ST rejected the idea that mindreading involves these kinds of principles, and they had several reasons for doing so. In the previous chapter we already encountered a very practical problem for TT: its inability to account for the context-sensitivity of our mindreading skills. Alvin Goldman (1989, pp.166-7) provides us with three other shortcomings: (i) TT-attempts to articulate the putative laws or 'platitudes' that comprise our folk theory are notably weak, (ii) this is strange when at the same time it is maintained that we constantly appeal to them in our understanding of others, and (iii) it remains doubtful whether children (at the age of 4-6) are sophisticated enough to employ these principles in the first place.

According to Goldman (1989), mindreading is 'process driven' rather than 'theory driven'. We are capable of accurately *simulating* a 'target system' (another human being) even if we lack a theory, as long as our initial mental states are the same as those of the target system and 'the *process* that drives the simulation is the same as (or relevantly similar to) the process that drives the system [that is, our own system]' (p.173). The idea that such a system of processes can be operated 'off-line' is integral to Goldman's version of ST. Robert Gordon (1992), in contrast, regards this as an 'ancillary hypothesis', though a 'very plausible one' (p.87). Gordon articulates a notion of *radical simulation* that involves a *transformation* at the personal level. Using our imagination, we are able to simulate what other persons think and feel and thus how they would behave, in their situation. However, we do not imagine *ourselves* in their situation; we imagine *them* in their situations by imaginatively occupying their situation. In some respects Gordon's notion of simulation resembles that of Jane Heal. Like Gordon, Heal (1986) stresses the importance of simulation as a transformation at the personal level: 'I place myself in what I take to be [the agent's] initial state by imagining the world as it would appear from his point of view and

[17] An interesting side-effect of the simulation movement is that it seems to pull the rug out from under eliminative materialism. As we saw in the previous chapter, eliminative materialism claims that there are no beliefs and desires because folk psychology is a radically false theory. But ST claims that the theory that posits a tacitly known folk psychological theory is *itself* radically false (cf. Gordon 1986, p.170; Goldman 1989, p.182).

then I deliberate, reason and reflect to see what decision emerges' (p.137). This is what she calls 'co-cognition', which is 'just a fancy name for the everyday notion of thinking about the same subject-matter [...] Those who co-cognize exercise the same underlying multifaceted ability to deal with some subject matter' (1998, p.483).

This chapter aims to determine whether the various ideas about ST articulated by the philosophers mentioned above offer a promising approach to intersubjectivity. First, I investigate the extent to which ST succeeds in providing a satisfying explanation of mindreading, understood as a functional process of mental state attribution (section 2). Next, I turn to versions of ST that try to go beyond mindreading by inserting simulation at a deeper level of intersubjectivity (section 3). Both attempts are accompanied by a number of problems, including some old ones (from the previous chapter) plus some new ones as well. I proceed by reviewing a relevant selection of the empirical evidence that is claimed to support ST, addressing various associated conceptual problems as I go (section 4). The chapter concludes by highlighting what I take to be the major 'internal' problems of ST - the problems that arise when one accepts a ST picture of intersubjectivity - and a more general comparison with TT (section 5).

2.1 Making sense of simulation

Simulation theory according to Goldman

Although advocates of ST reject the claim that mindreading is *theory-driven*, many of them remain surprisingly loyal to the idea that mindreading is primarily about the prediction and explanation of behavior according to the guidelines of belief-desire psychology. Goldman is an excellent representative of this line of thinking (especially in his earlier work), and his cognitivist version of ST is one of the more dominant players in the field.

According to Goldman, mindreading depends on a simulation process that involves the (introspective) use of the imagination and the attribution of 'pretend' mental states. Over the years, he has developed a full-blown heavyweight simulation system to explain what this means and how this works. The system is powered centrally by an impressive decision-making mechanism (see fig. 2.1). Goldman (2006) tells us that 'normally, our decision mechanism takes genuine (non-pretend) beliefs and desires as inputs and then outputs a genuine (non-pretend) decision. In simulation exercises, the decision mechanism is applied to pretend desires and beliefs and outputs pretend decisions' (p.29).[18] These pretend beliefs and desires express the idea that the attributor puts himself in the other agent's 'mental shoes', and they are fed into the decision-making mechanism when it is taken 'offline'. This results in what Stich and Nichols (1997) call 'pretense-driven offline simulation'.

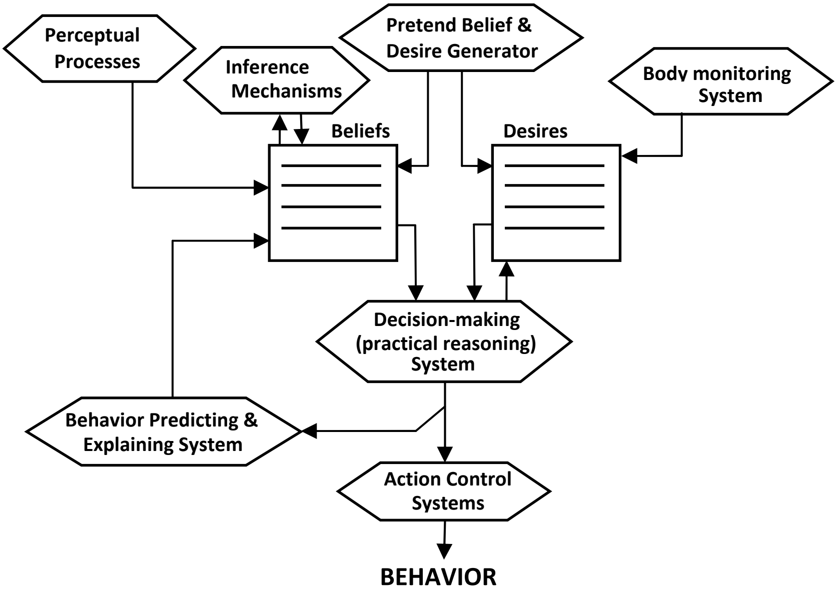

Perceptual Processes
Inference Mechanisms
Pretend Belief & Desire Generator
Body monitoring System
Beliefs
Desires
Decision-making (practical reasoning) System
Behavior Predicting & Explaining System
Action Control Systems
BEHAVIOR

Fig. 2.1 Off-line simulation account of behavior prediction (Nichols and Stich 2000)

Goldman (2006) proposes that simulations are structured as follows: 'First, the attributor creates in herself pretend states intended to match those of the target. In other words, the attributor attempts to put herself in the target's "mental shoes". The second step is to feed

[18] See Currie (1995) for a similar idea. Currie claims that in simulating another agent 'we tend to acquire, in imagination, the beliefs and desires an agent would most likely have in that situation, and those imaginary beliefs and desires have consequences in the shape of further pretend beliefs and desires as well as pretend decisions that mimic the beliefs, desires and decisions that follow the real case' (p.158).

these initial pretend states into some mechanism of the attributor's own psychology […] and allow that mechanism to operate on the pretend states so as to generate one or more new states [e.g., decisions] Third, the attributor assigns the output state to the target' (pp.80-1).

Such a functional procedure introduces a number of extra system requirements. In the first place, the mindreader projects pretend mental states onto the other agent on the basis of an *analogy* - because he knows how these mental states and behaviors are related in his own case. In order to do so, not only does he have to take his decision mechanism off-line in order to create pretend states, but he must also be able to reliably *identify* and *self-attribute* these mental states. The latter ability in turn requires a prior knowledge of the mental states in question. And even this does not guarantee a successful simulation, for there has to be a match in terms of a substantial resemblance between the attributed pretend state and its counterpart target state as well. Thus, Goldman's simulation procedure also requires a so-called 'resemblance model of other-attribution'. Together, these elements add a lot of philosophical baggage that requires inspection.

Some initial complications

It is important to notice that Goldman's simulation procedure heavily relies on the *argument from analogy*. According to the argument from analogy, we are able to infer that the bodily behavior of others is related to their mental states, since we have an intimate knowledge of our own mental states and their relation to our own bodily behavior. However, there are numerous problems with this argument.

Gilbert Ryle (1949) already claimed that it is a mistake to think that 'the spectator or reader, in following what is done or written, is making analogical inferences from internal processes of his own to corresponding internal processes in the author of the actions or writings. Nor need he [...] imaginatively represent himself as being, in the shoes, the situation and the skin of the author. He is merely thinking about what the author is doing along the same lines as the author is thinking about what he is doing, save that the spectator is finding what the author is inventing' (p.55).[19] Ryle also argued against the idea

[19] Interestingly, this comes close to Heal's description of 'co-cognition' - the ability to think about the same subject-matter. For Ryle, however, this process does not necessarily involve simulation.

of imputing to a variety of others what is true of my own simulated action, since this ignores the diversity of their actions. ‘The observed appearances and actions of people differ very markedly, so the imputation to them of inner processes closely matching [one's own or] one another would be actually contrary to the evidence’ (p.54).

Max Scheler (1973) raises a similar objection to the argument from analogy. He argues that when I infer or project the result of my own simulation onto your mind, I understand only *myself* in the situation - I don't understand *you*. Scheler’s work offers us various other objections against the argument from analogy as well.[20] For example, the argument from analogy is *developmentally unsound*, because the ability to infer or project on the basis of analogy is too difficult for young children, who are nevertheless capable of understanding others.

An important prerequisite for the analogy-based attribution of (pretend) mental states to others is self-attribution. Goldman (2006) remarks that this has been a serious problem for TT. Consider Nichols and Stich’s (2003) account of self-attribution, for example. According to this account, ‘to have beliefs about one’s own beliefs, all that is required is that there be a Monitoring Mechanism (MM) that, when activated, takes the representation *p* in the Belief Box as input and produces the representation *I believe that p* as output. This mechanism would be trivial to implement. To produce representations of one’s own beliefs, the Monitoring Mechanism merely has to copy representations from the Belief Box, embed the copies in a representation schema of the form *I believe that*___, and then place the new representations back in the Belief Box. The proposed mechanism (or perhaps a distinct but entirely parallel mechanism) would work in much the same way to produce representations of one’s own desires, intentions, and imaginings’ (Nichols and Stich 2003, pp.160-1; see also figure 1.1, chapter 1.3).

The problem with this account, according to Goldman (2006), is the fact that it leaves completely unanswered the question of how the Monitoring Mechanism decides which attitude *type* a targeted mental state belongs to. Is it a belief, a desire, or perhaps an intention? The problem is how the Monitoring Mechanism is able to determine that a given piece of mental syntax has this or that functional role. The traditional ‘solution’ of TT to the problem of self-ascription (which Goldman rejects) has been to assume that just *being* in a mental state *automatically* triggers a classification of yourself as being in that state (cf.

[20] See also Scheler (1973, pp.232-4), and Zahavi (2001, p.152) for an excellent summary and discussion of these objections.

Goldman 1993). But Nichols and Stich’s MM proposal is not really an improvement on this non-solution, since it also assumes that just *being* in a state of belief (or another propositional attitude) *automatically* triggers a classification of yourself as being in this state. The only difference is that it posits the *redeployment* or reuse of 'a piece of mental syntax’, namely the representation *p*.

But does Goldman’s own account fare much better? Goldman (2006) proposes that the first step towards identifying our own mental states involves a kind of ‘inner recognition’, which has to be understood as a perceptual process. Recognition is used in typing the target state, whether it’s a contentful or noncontentful state. Recognition is also used for classifying the target state in terms of supplementary features like strength or intensity. When we have identified our mental states as being contentful, they are either *redeployed*, or, when their format is ‘inadmissible’ (for example, in case of visual representations) they have to be *translated* into the right format: ‘For contentful target states, introspection uses either redeployment or translation to produce the content assignment contained in the metarepresentation’ (p.255). Thus, Goldman’s introspective model of self-attribution depends on three processes: recognition, redeployment and translation. But there is yet another requirement. In order to reliably identify and self-attribute mental states, the attributor must already have some understanding of them. As Goldman (1989) himself remarks, when an interpreter uses simulation to attribute mental states to another agent, this ‘assumes a prior understanding of what state it is that the interpreter attributes to [the agent]’ (p.182). And he insists that the meaning of these mental states is at least partly determined by their introspective properties.[21] At the same time, however, he readily admits that he lacks a satisfactory theory about how this works (cf. Goldman 2006, p.272).

It is not hard to see that Goldman’s story about introspection is philosophically very demanding. Now this is not necessarily a problem, as long as it gives us a satisfactory explanation of the phenomenon under consideration. But Goldman’s model seems to raise more questions than it answers. The processes it postulates are taken for *granted* (under the assumption that we need them in order to get the argument from analogy up and running), rather than properly *explained* (for example, in terms of their embodiment or

[21] He claims that ‘if the Simulation Theory is right […] it looks as if the main elements of the grasp of mental concepts must be located in the first-person sphere' (p.183). See also Goldman (2000), where he argues that he still subscribes to ‘a first-person, introspective understanding of mental state concepts’ (p.182).

development). And we have to add this to the fact that the argument from analogy is already problematic by *itself*. But there are other questions as well.

The argument from phenomenology revisited

According to Goldman, simulation is the primary and pervasive way of how we understand others. He claims that 'the strongest form of ST would say that all cases of (third-person) mentalization employ simulation. A moderate version would say, for example, that simulation is the *default* method of mentalization […] I am attracted to the moderate version […] Simulation is the primitive, root form of interpersonal mentalization' (2002, pp.7-8).[22] If this were true, then many of our everyday social encounters would involve complicated introspective processes, and we would be very busy creating and manipulating our pretend mental states, inferring and projecting them while hoping that they would match with those of the persons we try to understand. The question is whether this does justice to how we *experience* our daily meetings with other minds.

This is precisely the thrust of Gallagher's 'simple phenomenological argument'. Gallagher argues that if the simulation procedures prescribed by Goldman are explicit and pervasive, then we should be aware of the different steps that we go through as we consciously simulate the other's mental states. However, when I interact with others and try to understand them, 'there is no experiential evidence that I use such conscious (imaginative, introspective) simulation routines' (2007, p.65).

For simulation theorists, the easiest way to avoid the argument from phenomenology is to claim that we do not employ simulation routines in a *conscious* and *explicit* way during our social engagements. If simulation is an *unconscious and implicit* process, then what we experience or seemingly experience is not a good guide for what is 'really' happening in such cases, and the appeal to phenomenology would be inappropriate. As we saw in the previous chapter, this is a popular move for theory theorists, and as we will see in this chapter, many simulation theorists, including Goldman, pursue such a strategy as well.

There is another option, however. Instead of surrendering the personal level of social understanding so easily, one could bite the phenomenological bullet and reply that *there is*

[22] Goldman (1986) admits that in many cases, interpreters rely solely on 'inductively acquired information', but still this information is 'historically derived from earlier simulations' (p.176).

in fact experiential evidence that we use simulation routines in our social interactions. In his early work Goldman (1989) seemed to follow this line of argument, when he claimed that 'introspectively, it seems as if we often try to predict others' behavior - or predict their (mental) choices - by imagining ourselves in their shoes and determining what we would choose to do' (p.169). Paradoxically, however, at the same time he was also aware that the appeal to introspection could be used as a two-edged sword: 'There is a straightforward challenge to the psychological plausibility of the simulation approach. It is far from obvious, introspectively, that we regularly place ourselves in another person's shoes, and vividly envision what we would do in his circumstances' (p.176). But this didn't stop him from flirting with the idea that reliable self-attribution could be based on the phenomenological qualities of those mental states that are accessible to introspection. Goldman (1993), for example, proposed a 'sensible form of introspectionism', one that blocks introspective access to 'causal connections' but leaves open that people have 'introspective access to the mere occurrence of certain types of mental events' (p.373).

In his later work, however, Goldman becomes much more pessimistic about the prospects of phenomenological properties as suitable candidates for his introspective model of self-attribution (cf. 2006, p.249). Phenomenological properties are elusive, 'incapable of supporting weighty thesis', hard to agree upon and 'hotly disputed'. Goldman now argues that *neural properties* are 'natural candidates' for the input to introspective part of simulation. 'No challenge can be raised to their causal efficacy, and their detectability would be the same whether they were the substrate of conscious or of non-conscious mental states' (p.251).[23]

That the phenomenology of everyday social interaction is elusive and difficult to define or describe is also recognized by Gallagher (2004), who admits that introspective reports are 'notoriously suspect guides to what subjects are doing even at the conscious level' (p.94). Therefore, Gallagher thinks that an appeal to our social phenomenology should go *beyond* an appeal to good old introspection - to subjective reports about our everyday social encounters. Instead, he proposes to use phenomenology in its technical (Husserlian) sense, that is, as a strict method for the analysis of the common structures of experience. Phenomenology, thus understood, could be a promising research paradigm (cf. Gallagher and Varela 2003, Gallagher and Brøsted Sørensen 2006). For my current purposes, however, it goes too far to discuss its merits and limitations. What is important is

[23] The very idea of introspecting neural properties is briefly discussed in chapter 3.3.

that the simple phenomenological argument *by itself* is sufficient to counter an explicit ST approach to intersubjectivity.

Goldman (2006) has attempted to circumvent possible phenomenological objections such as the phenomenological argument by claiming that a great deal of simulation is semi-automatic, non-conscious or minimally conscious. He now proposes a distinction between *low-level* and *high-level* simulation. High-level simulation involves the *conscious* use of our imagination to manipulate propositional attitudes such as beliefs and desires, whereas low-level simulation is 'simple, primitive, automatic, and largely *below the level of consciousness*' (p.113, italics added). High-level simulation is distinct from low-level simulation in that it includes 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 simulation routine are subject to voluntary control; and (c) the process has some degree of accessibility to consciousness. However, since the simple phenomenological argument is directly aimed at criterion (c), it could be argued that high-level simulation is still vulnerable to Gallagher's criticism.

Goldman does have some elbow room, however. For example, he could further downplay the importance of introspective access for high-level simulation, since both his recognitional model of self-attribution and his resemblance model of other-attribution are already fueled by neural instead of phenomenological properties. Also, he could further downplay the importance of high-level simulation *itself*, emphasizing instead the crucial role of low-level or 'tacit' simulation for our meetings with other minds. And finally, he could point out that, when it comes to the question of phenomenology, ST is no worse off than its competitors. Goldman (1995), for example, already claimed that 'it is a psychological commonplace that highly developed skills become automatized, and there is no reason why interpersonal simulation should not share this characteristic (On the issue of conscious awareness, the ST is no worse off than its competitors. Neither the rationality approach nor the folk-TT is at all credible if it claims that appeals to its putative principles are introspectively prominent aspects of interpretation)' (p.88).

However, all these options force ST to abandon the *personal* level of description. The simple phenomenological argument again seems to be strong enough to drive a wedge between claims about our *conscious* experience of social understanding on the one hand, and claims about the mechanisms and processes that *unconsciously* facilitate such an understanding on the other hand. It can be used to cast doubt on ST insofar the latter postulates complicated introspective procedures and the explicit manipulation and attribution of mental states. But of course ST is not necessarily committed to all these heavy assumptions. Besides looking for evidence on the sub-personal level, simulation theorists could also try losing weight by discarding some of the cumbersome personal level assumptions. Instead of explaining in terms of simulation what is, in essence, a very *narrow* conception of intersubjectivity as mindreading, one might as well try to use the notion of simulation to *broaden* its scope. This is where Gordon's 'radical' simulation comes in.

Simulation theory according to Gordon

According to Gordon, simulation proceeds by exercising a skill that has two components: the capacity for practical reasoning - roughly, for making decisions on the basis of facts and values - and the capacity to introduce 'pretend' facts and values into one's decision making (which is typically done to adjust for relevant differences in situation and past behavior). When we simulate others, we predict what they will decide to do by making a decision ourselves: a 'pretend' decision, which is made in our imagination and with adjustments for the relevant differences. Gordon (1986) describes this process as follows: 'Our decision-making or practical reasoning system gets partially disengaged from its "natural" inputs and fed instead with suppositions and images (or their "subpersonal" or "sub-doxastic" counterparts). Given these artificial pretend inputs the system then "makes up its mind" what to do. Since the system is being run off-line, as it were, disengaged also from its natural output systems, its "decision" isn't actually executed but rather ends up as an anticipation [...] of the other's behavior' (p.170). Where Goldman gives pride of place to the capacity to *explain* or *interpret* the behavior of others in terms of mental states, Gordon focuses mainly on the role of simulation in *prediction* or *anticipation*. But there are other differences as well.

Gordon's radical simulation is radical in the sense that it inserts simulation at a deeper level of intersubjectivity. Simulation is not simply part of a matching process between mental states - a mere cognitive *heuristic*, as it is for Goldman. Rather, it allows us to recognize the other as 'mind-endowed' in the first place (Gordon 2004, p.2). Radical simulation can be considered as a 'lightweight' version of ST, because Gordon distances himself from three elements that are involved in mindreading ST accounts: (i) an analogical inference from oneself to others; (ii) premised on introspectively based attributions of mental states to oneself; (iii) requiring prior possession of the concepts of the mental states ascribed (cf. Gordon 1995, p.53).

According to Goldman's heavyweight version of ST, I set out to predict someone's decision by imagining myself in her mental shoes. In order to do this, I have to create a pretend decision, introspect this decision and 'transfer' it to her. I do this on the basis of an analogical inference – that she is 'like me'. But Gordon thinks that this is problematic. He argues that, when I simulate someone, I do not imagine *myself* in her situation. Instead, I try to imagine *the other* in her situation by imaginatively occupying her situation. This involves a personal-level 'transformation' of myself into her, an 'egocentric shift', or a 'recentering' of the egocentric map. No further mental state management is required. 'The point I am making is that once a personal transformation has been accomplished, there is no remaining task of mentally transferring a state from one person to another, no question of comparing [the other person] to myself. For insofar as I have recentered my egocentric map on [the other person], I am not considering what [I] would do, think, want, and feel in the situation' (Gordon 1995, p.54). When I recenter my egocentric map on you, I do not consider what *I* would think, want or decide; instead, I imagine, in the first-person, how *you* see the world.

The central idea behind this form of 'actual simulation', as Stich and Nichols (1997) have termed it, is that what are essentially *first-person* decision procedures can be applied to others by transforming ourselves into other 'first persons'. Gordon (1995) argues that the method we ordinarily use is limited to identifying states in the first person, but, thanks to our capacity for imaginatively transforming ourselves into other 'first persons', it is not exclusively a one-person method.

Simulation, thus understood, frees me from the task of making analogical inferences from me to you. Moreover, it is also devoid of any conceptual wizardry since I am not concerned with mental states at all. This allows Gordon to evade an argument he himself launched against TT, namely that it demands 'a highly developed theoretical intellect and a methodological sophistication rivaling that of modern-day cognitive scientists. That is an awful lot to impute to the four-year-old, or to our savage ancestors' (1986, p.71).

Goldman's version of ST holds that the attributor has to make an *introspective identification* of his pretend decision in order to project it onto the target. Gordon, however, rejects this element as well. Instead, he offers an interesting alternative: *ascent routines*. Suppose you are asked whether you believe it is raining. On Goldman's simulation model, the canonical way to answer questions of this type is to look inwards in order to inspect the phenomenological qualities of the belief state that it is raining outside. Gordon, however, denies that you have to do this. He suggests that, instead, you simply have to ask yourself: 'Is it raining outside?' If the answer is 'Yes', then you report that you believe it is raining outside. Gordon adopts this idea from Gareth Evans (1982), who proposed that we can encapsulate the procedure for answering questions about what one believes in the following simple rule: whenever you are in a position to assert that p, you are ipso facto in a position to assert 'I believe that p'. Evans argues that we answer questions about our own beliefs by using a redeployment strategy: 'I get myself in a position to answer the question whether I believe that p by putting into operation whatever procedure I have for answering the question whether p' (p.225).

What is important about ascent routines, according to Gordon (2007), is not so much the question-answer form, but the fact that, whether in answer to a question or not, people optionally step up a semantic level from an assertion that p to a self-ascription of a belief that p. By doing this, we move from an expression of the belief that p to a self-ascription of the belief that p. 'Thus, we may move from an assertion about the weather, "It's raining," to an assertion about ourselves, "I believe it's raining," from a weather report to a self-report. The permissibility of this move from asserting that p to affirming that one believes that p is closely related to the impermissibility of asserting that p and denying that one believes that p' (p.154).

Although this explains how we step up from an assertion to a self-ascription of a belief, it only does so for our *own* case. In order to ascribe beliefs to others, according to Gordon, ascent routines need to be embedded in *simulations*. For example, I want to know whether *someone else* believes it is raining. First, I have to transform myself into the other by imaginatively occupying his situation. This involves an 'egocentric shift' or a 'recentering of the egocentric map'. Second, I ask myself, in the role of the other, the question 'Is it raining?' and my simulation links the answer to the particular individual whose situation and behavior constitute the evidence on which the simulation is based - the individual whom one is identifying with within the simulation. If the answer is affirmative, I can make the assertion 'He believes it is raining'. Thus, Gordon (1995) argues, 'to ascribe to O a belief that p is to assert that p within the context of a simulation of O' (p.60).

Compared to Goldman’s proposal, Gordon’s description of ascent routines gives us a much more parsimonious account of self and other ascription, in the sense that it radically discounts the importance of introspection, analogical inference and mental state management. At the same time, however, it remains somewhat mysterious how we should think of simulation as a *transformation* (an ‘egocentric shift’) at the personal level. Gallagher (2007) remarks that ‘although Gordon does away with the need for an extra step involving inference, because we are “already there” in the other’s perspective, these transformations still require an “as if” component. Otherwise, my own first-person perspective on the world would simply collapse into the first-person perspective of the other and the self/nonself distinction would disappear’ (p.67). He argues that this makes radical simulation, understood as a *personal level* transformation, an easy target for the simple phenomenological argument, since neither the ‘as if’ component, nor a collapse of the self/nonself distinction are part of our everyday social experience.

In most second-person engagements, according to Gallagher, there are all kinds of contextual constraints that help us to differentiate between our own first person perspective and that of others. ‘When I look out of the window and see a man standing across the road I don't have to transform myself into his perspective to know that he happens to see the road from an angle that differs from my view. I can see that this must be the case simply from the differences that define our positions vis-à-vis the road, and from the orientation and postural stance of his body’ (p.68). If these contextual constraints prevent us from understanding the man’s behavior (for example, his sudden burst of excitement), we do not so much attempt to transform ourselves into him, but rather try to move to a position similar to his in order to see what he is seeing. This is not so much simulation, but actual physical movement. Gallagher admits that this is of course not always possible. Our options for physical movement could be limited, for example, or there could be other severe constraints that prevent us from understanding what the man is excited about. When this happens, according to Gallagher, we could try to put ourselves in the other’s shoes. However, even in these cases it is still not clear how a simulation would yield the right explanation of his behavior: 'Without further information, simply by transforming my egocentric perspective into his I will remain puzzled. Perhaps, by simulation, I would hypothesize that he is playing a joke on me, or, by appeal to theory, that he is delusional. But I would still need more information about the man’s character - I would need to know the man’s story – to determine whether my simulative [...] supposition was correct’ (ibid.).

I agree with Gallagher that Gordon’s idea of an imaginative transformation at the personal level *by itself* is not sufficient to explain our understanding of others. What is also needed is an explanation of how we acquire the necessary background knowledge about other people and the various pragmatic contexts in which we encounter them. This is necessary in order to ensure that my imaginative transformation meets the demands of context-sensitivity, i.e. incorporates adjustments for the relevant differences. But I don’t see why the ‘as if’ component that is characteristic for such a transformation would be very problematic. In fact, I think that here the appeal to phenomenology actually works *against* Gallagher. Sometimes, we do experience an ‘as if’ component when we try to put ourselves in the other’s shoes, and sometimes, we are perhaps not as sure about the self/other distinction as we would like to be. At the same time, however, Gallagher is certainly right that this is not our *default* position.

If we grant Gordon that we sometimes try to understand others by imaginatively occupying their situation (in a non-mentalistic way), then the question is how we can *explain* this social ability. Gordon is not very clear about this. He claims that simulation involves the interpretation of the behavior of others under the ‘same scheme’ that makes our own behavior ‘intelligible’ to us. This requires a basic understanding of the ‘intentional scheme of reasons and purposes’, one that directly engages ‘productive processes such as practical reasoning, emotion formation and decision making’ (Gordon 2005, p.101). And this kind of understanding is meant to play a vital developmental role, for the ‘implicit recognition is crucial to understanding how we bootstrap ourselves into an explicit folk psychology. Bootstrapping is possible because intentional explanations in terms of reasons, purposes and objects are at least implicitly mental’ (p.105). Gordon’s emphasis on the implicitness of this kind of mental recognition seems to suggest that we will not find evidence for it on the personal level. But if we are supposed to descend to the level of sub-personal processes, then it is not clear what is meant by ‘reasons, purposes and objects’ that are ‘implicitly mental’. Moreover, the question is whether these sub-personal processes are best characterized in terms of *simulation.*

Simulation theory according to Heal

Although Heal’s ideas about simulation are somewhat different from those of Gordon, she also stresses the importance of a transformation at the personal level: ‘I place myself in what I take to be [the agent’s] initial state by imagining the world as it would appear from his point of view and then I deliberate, reason and reflect to see what decision emerges’ (1986, p.137). She even calls this an ‘a priori truth’, and claims that ‘thinking about others’ thoughts *requires* us, in usual and central cases, to think about the states of affairs which are the subject matter of those thoughts, i.e., to co-cognize with the person whose thoughts we seek to grasp’ (1998, p.484; italics added).

Heal distinguishes her claim from the contrasting claim (defended by Goldman) that, when we think about other's thoughts, we sometimes ‘unhook’ our cognitive mechanisms so that they can run ‘off-line’, and then feed them with 'pretend' versions of the sorts of thought we attribute to the other. She argues that the first claim, about the importance of simulation as co-cognition, should be the focus of the ST debate. The second claim is nothing more than an empirical hypothesis about the way co-cognition is realized. It can be refuted, but if that happens, is does not necessarily undermine the first claim, since there may be other ways of realizing co-cognition.

Heal’s notion of co-cognition is different from Gordon’s notion of radical simulation in the sense that it only seeks to illuminate how we predict the thoughts of others in cases where we *already have* information about their background beliefs and desires. She gives the following example: ‘Suppose I wish to predict what John will think of the new jacket; will he think it garish? Suppose further that I know that John believes the jacket to be scarlet and he thinks all bright colors to be garish. I will, of course, expect him to think the jacket garish’ (1995, p.39). In cases such as this one, according to Heal, we co-cognize with others by harnessing our own cognitive apparatus and making it work in parallel with that of the other. Given the presupposition that we already are in the possession of the background knowledge required to interpret others in a context-sensitive way, it seems hard to disagree with Heal’s modest proposal that thinking about others requires us to think about the same subject matter. At the same time, however, the more interesting question of *how* we acquire this background knowledge remains unanswered.

Another important difference with Gordon is that Heal argues that the ability to engage in co-cognition and draw conclusions about what another is thinking presupposes the mastery of mental concepts. She remarks that the output of a simulation of another's thought processes is in fact a *judgment* that someone else is having a thought of a certain sort. This means that one must already have the *concept* of belief in order to *simulate* the belief that *p* (cf. Heal 1995). Of course, what is required here is a story about mental concept acquisition. But there is another important requirement as well. According to Heal, the conclusions we draw about the thought processes of other agents can only be justified on the assumption that they are, at least in a very minimal sense, *rational* agents like us. Given the assumption of such a minimal form of rationality, Heal attempts to show why reliance on co-cognition seems to be a sensible way to proceed in trying to grasp where another's reflections may lead. 'The other thinks that p1 – pn and is wondering whether q. I would like to know what she will conclude. So I ask myself "Would the obtaining of p1 – pn necessitate or make likely the obtaining of q?" To answer this question I must myself think about the states of affairs in question, as the other is also doing, i.e. I must co-cognize with the other. If I come to the answer that a state of affairs in which p1 – pn would necessitate or make likely that q, then I shall expect the other to arrive at the belief that q' (1998, p.487).

Although co-cognition is put forward as a species of simulation, it is very much dependent on certain normative principles of *rationality* in order to get off the ground. We can only make sense of others and co-cognize with them on the assumption that rationality imposes certain requirements, or normative rules, on what they think and how they behave. In this respect, Heal's version of ST is strongly committed to rationality theory, or 'normative TT'. Rationality theory (RT) is most prominently defended by Davidson (1984) and Dennett (1987) as an account of intersubjective interpretation. The core idea is that interpretation proceeds by making the charitable assumption that others usually comply with certain normative principles of rationality: for example, that rational agents believe truths, their belief-sets are more or less coherent, and their desires are aimed at things that is good for them to have (cf. Goldman 2000). According to RT, these principles of rationality guide the process of mindreading in roughly the same way as the theoretical generalizations postulated by TT.

Whether or not RT is problematic mainly depends on how the notion of rationality is unpacked. If rationality is defined in a very strict sense, e.g. as a firm understanding of the rules of logic, then RT is not very plausible as an account of everyday intersubjective interpretation.[24] But if the notion of a rational agent becomes so vague and empty that is can be replaced by something like 'any typical person' (cf. Perner 1996, p.92), then it loses all its explanatory power. This poses a potential difficulty for Heal, at least insofar her account of co-cognition relies on the assumption of minimal rationality. I certainly do not want to deny that something like co-cognition is indispensable if we want to *think about the thoughts of others* (although this is just one aspect of intersubjective understanding). At the same time, however, I do not really see how Heal's appeal to simulation provides us with a satisfying *explanation* of this ability.

A threat of collapse and the return of folk psychological principles

One of the most important problems for Goldman's version of ST is its inability to account for the *context-sensitivity* of our intersubjective understanding. To understand why this is so, we have to recall that ST needs to explain how mindreading can be exercised for the purposes of both behavior prediction and explanation. If we use simulation for behavior prediction ('forward' simulation), we feed hypothetical beliefs and desires into our own off-line decision mechanism and we predict what the agent would decide to do, given those beliefs and desires. As Gallagher (2007) notices, this is not unproblematic since it presupposes that we already have some idea what is going on with the other person. 'Where does that knowledge come from and why isn't that already the very thing we are trying to explain?' (p.64). But there may be even more serious problems when it comes to using simulation for behavior *explanation* ('backward' simulation). Proponents of ST à la Goldman often suggest that this requires something akin to a 'generate-and-test' strategy:

[24] This has to do with the questionable grasp of logic by ordinary people, let alone children. The latter already show substantial mastery of attribution skills in their attitude ascriptions. According to RT, then, these children must understand the rules of logic. But it is really plausible to suppose that they grasp the general notions of logical consistency and deductive closure? Actually, it is doubtful whether even untrained adults grasp these notions. Many scientific studies of deductive reasoning challenge the notion that untrained adults approach such tasks with abstract semantical or proof-theoretic concepts of the sorts used in formal logic (Cheng and Holyoak 1985, Cosmides 1989). Similarly, psychological studies of decision and choice challenge the notion that naive people utilize standard normative models (Tversky and Kahneman 1986).

we try to find the right beliefs and desires which, when fed into our off-line decision mechanism, will produce a decision to perform the behavior we want to explain.[25]

However, the problem is that there are *far too many* hypothetical beliefs and desires that lead to the behavior in question. Although sometimes certain belief-desire pairs are easily excluded on the basis of information about the agent's perceptual situation or pre-existing knowledge of the agent's beliefs and desires (but how do we acquire this?), it will often be the case that there are lots of alternative explanations that can't be excluded in this way. According to Goldman (1989, pp.178-91), in these cases we simply have to assume that the agent is psychologically similar to us, attribute beliefs that are 'natural for us' and reject (or perhaps do not even consider) hypotheses attributing beliefs that we consider to be less natural. Gordon (1986) tells a similar story: 'No matter how long I go on testing hypotheses, I will not have tried out all candidate explanations of the [agent's] behavior. Perhaps some of the unexamined candidates would have done at least as well as the one I settle for, if I settle perhaps indefinitely many of them would have. But these would be "far fetched", I say intuitively. Therein I exhibit my inertial bias. The less "fetching" (or "stretching", as actors say) I have to do to track the other's behavior, the better. I tend to *feign* only when necessary, only when something in the other's behavior doesn't fit. This inertial bias may be thought of as a "least effort" principle: the "principle of least pretending". It explains why, other things being equal, I will prefer the less radical departure from the "real" world -i.e. from what I myself take to be the world' (p.164).

While this seems to be an attractive and parsimonious proposal, the question is how to explain the fact that we often *do* make rather impressive adjustments in our understanding of other agents. Remark that what is at issue here is basically the same

[25] Goldman (2006) explains this as follows: 'In decision prediction, the target's initially specified states are presumptive causes of a subsequent effect or outcome, which is to be calculated. The mindreader moves 'forward' from the prior evidence events to their effect. Many mental attributions, however, must fit a second pattern, in which a sought-after mental state is the cause of some known (or believed) effects. Here the attributor moves 'backward' from evidence states (observed behavior, facial expressions, etc.) to the mental cause of interest [...] This type of mindreading might be approached via a generate-and-test strategy. The attributor begins with a known effect of a sought-after state, often an observable piece of behavior. He generates one of more hypotheses about the prior mental state or combination of states that might be responsible for this effect. He then 'tests' (one or more of) these hypotheses by pretending to be in these states, feeding them into an appropriate psychological mechanism, and seeing whether the output matches the observed evidence. When a match is found (perhaps the first match, or the 'best' match), he attributes the hypothesized state or combination of states to the target' (p.45).

problem that bothered TT: how can we account for the context-sensitivity of our intersubjective skills? However, whereas TT approached this question from a third-person perspective, ST tries to answer it by taking the first-person perspective for granted. But how are we able to bridge the distance between our own beliefs and desires and those of agents who are very different from us? Since simulation does not provide us with the necessary resources to determine which beliefs and desires to put aside and which to keep in play, it is not at all clear how we end up having the appropriate ones and arrive at the right kind of understanding of others. Although Gordon (unlike Goldman) is not *per se* committed to an explanation of this ability in terms of (the reconstruction of) belief-desire pairs, he needs to say at least something about *how* it works. His ascent routine proposal could be a first step in the right direction, but this requires much more elaboration (cf. chapter 5.5).

Several TT proponents argue that this problem indicates that ST cannot give an adequate explanation of our intersubjective skills without appealing to theoretical principles. And some advocates of ST admit that this indeed appears to be the case. Goldman (2006), for example, agrees that simulation processes need theoretical backup: 'The generate-and-test strategy employs simulation at a crucial juncture but also relies on theorizing. Theorizing seems necessary to generate hypotheses about states responsible for the observed effects, hypotheses presumably prompted by background information. Thus, *pure simulationism is inapplicable here*' (p.45, italics added).[26]

There is yet another way of demonstrating that ST is in need of theory. Consider the following argument against ST made by Dennett (1987): 'An interesting idea [...] is that when we interpret others we do so not so much by *theorizing* about them as by *using ourselves as analog computers* that produce a result. Wanting to know more about your frame of mind, I somehow put myself in it, or as close to being in it as I can muster, and see what I thereupon think (want, do...). There is much that is puzzling about such an idea. How can it work without there being a kind of theorizing in the end? For the state I put myself in is not belief but make-believe belief. If I make believe I am a suspension bridge and wonder what I will do when the wind blows, what "comes to me" in my make-believe

[26] See also Goldman's statement that 'in a decision-prediction task, an attributor would use theoretical reasoning to infer the target's initial states (desires and beliefs), for which the corresponding pretend states are constructed. The pretend states are then fed into the decision making mechanism, which outputs a decision. The first step of this sequence features theorizing, whereas the remaining steps feature simulating' (2006, p.44).

state depends on how sophisticated my knowledge is of the physics and engineering of suspension bridges. Why should my making believe I have your beliefs be any different? In both cases, knowledge of the imitated object is needed to drive the make-believe "simulation," and the knowledge must be organized into something rather like a theory' (pp.100-1).

Goldman initially parried this argument by making a distinction between *theory-driven* and *process-driven* simulation. Process driven simulation does not collapse into theorizing, according to Goldman, as long as (i) the process driving the simulation of the other is the same as the process that drives our own system, and (ii) we start out with the same mental states. But in his later work he admits that this response has been too quick. For even if we think of simulation as being process-driven, such a process still requires that 'some elements inside the attributor causally mediate between his explicit premises and conclusions, and that the causal structure of these elements mirrors the logical structure of psychological theory' (2006, p.33). If this is true, then simulation depends on tacit theory. And this in turn raises the question whether and to which extent ST and TT are in fact *rivals*. Are both positions indeed as incompatible as they claim to be? Here it is interesting to consider Goldman's final observation with respect to the problem of collapse. He points out that, although there is a prima facie conflict between simulation and theory at the personal level, there is no conflict between them at different levels. 'There is nothing wrong in supposing that mindreading is executed at the personal level by simulation, which is in turn implemented at the sub-personal level by an underlying theory. Indeed, some might say, how could simulation be executed unless an algorithm for its execution is tacitly represented at some level in the brain? Isn't such an algorithm a sort of theory?' (ibid.). Now this is a very dangerous move. For Goldman left the personal level when he argued that simulation is to a large extent 'non-conscious or minimally conscious' and disqualified the phenomenology of intersubjectivity as notoriously unreliable. If, as a result, decisive evidence for ST has to be found on the *sub-personal* level, it is very strange to claim that this evidence could at the same time be interpreted as evidence for TT.

At this point, the only way out for ST seems to propose some sort of collaboration with TT and promote a 'hybrid treatment'. And this is precisely Goldman's strategy. Arguing that 'the generate-and-test strategy requires cooperation between simulating and theorizing', he adopts a mixed-method approach that accommodates both simulation and theorizing. However, this approach still emphasizes simulation as the default procedure. 'Our fundamental, default procedure is to project our own basic concepts and combinatorial principles onto others' (2006, pp.175-6). Although theoretical principles may be necessary for mindreading, their work is subservient and supplemental to that of simulation routines. But there are also hybrid theorists who see the roles of theory and simulation *reversed.* They hold that if simulation plays a vital role in our understanding of others, it does so by feeding the outputs of simulation routines into theorizing activities that brings folk psychological principles into play. Theory still does the heavy lifting in explaining the other's behavior (cf. Carruthers 1996).[27]

Hutto (2008a) notices that even those hybrid theorists who place less emphasis on the acquisition of folk psychological principles are still convinced that theory has to play *some* role in our intersubjective encounters. For example, Stueber (2006) claims that the 'competence in the full range of folk-psychological concepts that we normally attribute to adult human beings requires some minimal *theoretical grasp* of the nature of mental states and how they might interact [...] such a concession does not imply that folk-psychological concepts requires possession of a very rich theory that involves knowledge of detailed theoretical principles about the interaction of various mental states' (p.149, italics added).

One way or the other, the conclusion is that ST cannot solve the problem of context-sensitivity by itself. Insofar as it tries to explain intersubjectivity in terms of *mindreading*, it needs to be supported by (i) theoretical principles (belief-desire syllogisms) that structure our mental state attributions in terms of belief/desire pairs, and (ii) tacit theoretical knowledge in order to determine which belief-desire pair does the actual job of predicting/explaining the behavior under consideration. This, however, amounts to a *restatement* of all the TT problems mentioned in the previous chapter. These objections are obviously most acute for Goldman's version of ST. But Heal's account of co-cognition is vulnerable as well, since she is also committed to a 'principled' view of intersubjectivity.[28] Gordon seems to be the only one who radically rejects an appeal to theoretical or rational principles. At the same time, however, it is not clear how his own radical brand of ST accounts for the context-sensitive application of our intersubjective skills.

[27] The increasing number of hybrid ST/TT accounts makes it increasingly difficult to maintain a strict distinction between TT and ST, even with respect to their basic assumptions. For the many fine distinctions that have been drawn within the theory/simulation contrast and some challenges to the distinction itself, see Davies and Stone (1995a, 1995b).

[28] Although Heal's version of ST is committed to RT, in some respects it comes close to TT as well. For example, Heal (1994) grants TT that 'people who think about others' thoughts know such generalities as that beliefs and desires tend to lead to action' (pp.141-2).

2.2 Assessing the empirical evidence

Again, the false belief test

Many simulation theorists maintain that their arguments are supported by empirical evidence. We already encountered an important source of evidence from developmental studies in our discussion of TT: the false belief test. A good summary of the classic false belief test (Wimmer and Perner 1983) and its key result is given by Gordon (1986): ‘The puppet-child Maxi puts his chocolate in the box and goes out to play. While he is out, his mother transfers the chocolate to the cupboard. Where will Maxi look for the chocolate when he comes back? In the box, says the five year old, pointing to the miniature box on the puppet stage: a good prediction of a sort we ordinarily take for granted [...] But the child of three to four years has a different response: verbally or by pointing, the child indicates the cupboard. (That is, after all, where the chocolate is to be found, isn't it?) Suppose Maxi wants to mislead his gluttonous big brother to the *wrong* place, where will he lead him? The five year old indicates the cupboard, where (unbeknownst to Maxi) the chocolate actually is [...] The *younger* child indicates, incorrectly, the box’ (p.168).

Despite the fact that these results are often claimed to provide evidence for certain (internalist) versions of TT, Gordon (1986) claims that they actually show that there is something *wrong* with TT. For if TT is correct, Gordon argues, then children would not be able to predict or explain human action *prior* to the internalization of a folk psychological theory. But *after* the internalization of such a theory, they would be able to deal indifferently with both the actions caused by true beliefs and the actions caused by false beliefs. It is hard to see how the semantical question could be relevant in this respect. However, the finding that children *do* respond differentially to these actions is just what we should expect if ST is correct. ST predicts that, prior to developing the capacity to simulate others for purposes of prediction and explanation, children will make *egocentric errors* in predicting and explaining the actions of others. They will predict and explain as if whatever they themselves count as ‘fact’ were also fact to others. What the false belief test indicates, according to Gordon, is that children of three to four years are only capable of a kind of ‘first person pretend play’. They are able to simulate decision procedures in order to predict their *own* behavior in hypothetical situations, but fail to make ‘adjustments for the relevant differences’ when it comes to predicting the behavior of others. In these latter cases, they resort to ‘total projection’ (1986, p.162). Goldman (2006) suggests that we should understand this projection in terms of a ‘quarantine-violating simulation process’, in which the quarantine violation strongly affects the resulting attribution: ‘projection occurs when a genuine, nonpretend state of the attributor seeps into the simulation routine despite its inappropriateness (as judged by information the attributor possesses). This results in an attribution that is inappropriately influenced by the attributor’s own current states (genuine, non-pretend states)’ (p. 165).

However, it is not clear why the results of the false belief test would be incompatible with TT. Stich and Nichols (1992), for example, have argued that it is possible that children of three to four years have mastered *only part of a theory* that specifies how beliefs and desires lead to behavior: ‘at this stage, they might simply assume that beliefs are caused by the way the world is; they might adopt the strategy of attributing to everyone the very same belief they have. A child who has acquired this much of folk psychology would incorrectly attribute to Maxi the belief that the chocolate is in the cupboard’ (p.60). This is what they call ‘default’ attribution.

Furthermore, Harris (1992) has pointed out that, given the original motivation behind the false belief test, we should not expect it to be congenial to ST and problematic for TT. The initial popularity of the false belief test was due to the fact that it made it impossible for children to use a very simple strategy (such as a total projection or default attribution) in order to achieve predictive success (cf. chapter 1.4). Because such a strategy would not provide the appropriate evidence for the existence of a theory of mind, researchers started to use the false belief task because it required something more sophisticated. Now we might argue about whether this ‘something more’ should be interpreted as simulation or theory, but Harris’ point is that there is no reason to think in advance that the false belief test is likely to support ST over TT.

Before continuing, let us briefly consider the development of self and other attribution. Some advocates of ST (Goldman, for example) are committed to the view that we make analogical inferences about the other’s mental states on the basis of an introspective model of self-attribution. This presupposes that children attribute mental states to themselves before they attribute them to others. However, as we saw in the previous chapter, a number of experiments seem to indicate that self- and other-attribution develop in *tandem* (Gopnik and Wellman 1992, Gopnik and Meltzoff 1994). If this is true, then it poses a problem for those versions of ST that rely on the primacy of self-attribution. Nonetheless, the debate on this topic is all but decided.

Imitation and pretend play

Simulation theorists might also point to so-called ‘precursors’ to simulation. If intersubjectivity depends on the ability to simulate the thoughts, feelings and behaviors of others, these precursors could show us how this ability unfolds during development. *Imitation* might be such a precursor.

Numerous experiments indicate that young children have strong conventional and conformist tendencies. Meltzoff and Moore (1977, 1994), for example, demonstrated that neonates are able to pick out a human face from the crowd of objects in its environment and imitate the gesture it sees on that face. By 14 months, infants imitate a modeled novel act after a week’s delay (Meltzoff 1988, 2004; see also Gergely et al. 2002). And by 15-18 months, infants recognize the underlying goal of an unsuccessful act they see modelled, and re-enact it, using various means.

Imitative behavior does not disappear with age. On the contrary, adults continue to imitate and learn to copy increasingly complex patterns of behavior. This is known as the ‘chameleon effect’ (Chartrand and Bargh 1999), or, in the context of emotion-related behaviors, ‘emotional contagion’ (Hatfield et al. 1994). Human beings automatically tend to assimilate their behavior to their social environment, and react strongly to modelled or represented personality traits and stereotypes. Therefore, it has been suggested that imitation functions as a kind of ‘social glue’ that makes it easier for people to coordinate actions and interact in a smooth way (Dijksterhuis 2004, Chartrand and Bargh 1999).

Without doubt, these findings show that imitation is important to intersubjectivity. But imitation is still one step short of *simulation*. An important difference is that imitation does not require the ‘as if’ component, which is central to simulation. It is often suggested that the imitative tendencies of young children are due to a lack of inhibitory control. The idea is that their perception of behavior tends to be enacted automatically in imitative behavior, unless it is actively inhibited. As a result, they are not yet capable of pretending, of acting ‘as if’. Inhibition is a function of frontal areas of the brain, but babies and very young children do not yet have a well-developed frontal function or capacity to inhibit imitative tendencies (Kinsbourne 2004). It has been shown that adults with damage to certain frontal areas of the brain also imitate uninhibitedly (Lhermitte et al. 1986, Lhermitte 1986). Patients with this 'imitation syndrome' compulsively imitate gestures or even complex actions, although they have not been instructed to do so. Moreover, they keep on doing this even when this behavior is socially unacceptable or odd, such as putting on eyeglasses when one is already wearing glasses. The tendency to imitate is not confined to young children or patients with frontal lobe damage. While normal adults are usually able to inhibit overt imitation selectively, overt imitation can be seen as a surface symptom of non-stop inhibited imitation. Kinsbourne (2004) proposes that covert imitation may reflect a basic motivation of human beings to interact synchronously or entrain with one another, which is a mechanism of affiliation as well as of social perception and learning. This suggests that imitation is ontogenetically more basic than simulation, since the latter requires a certain amount of frontal lobe development to facilitate the 'as if' component. There is another subtle difference between imitation and simulation. Simulation can be defined in the sense of a simulator: a model that we can *use* so we can understand the real thing. But imitation is rather triggered by *others* than actively initiated by the self. This suggests that imitation also lacks the 'instrumentality condition' which is characteristic of simulation.[29]

It is interesting to contrast the above findings with Goldman's (2006) suggestion that inhibition plays a central role in enabling children to override their egocentric tendencies. Goldman thinks that inhibitory control is required to keep them from projecting their own characteristics onto others. According to him (and many other simulation theorists), total projection is the most basic form of simulation since it involves a total projection of one's own first person mental states (beliefs, desires etc.) onto others without adjusting for the relevant differences. However, the fact that inhibition is also required to override excessive *imitation* gives rise to the question to which extent these mental states can be said to be

[29] In this section I have used the term imitation in a rather broad sense. However, it is possible to give a more narrow definition of imitation, one that goes beyond a mere 'copying' of behavior and requires not only novelty but also a means/end structure. Such a definition might be able to incorporate the pretense and the instrumentality condition, and this would blur the distinction between simulation and imitation. But even in this case, imitation as the copying of behavior would be much more basic (at least from a developmental perspective) than imitation as the combining of behavioral means with intentional goals in a novel way. Moreover, this last notion still falls way short of the kind of simulation that is presupposed by ST, since it deals with the manipulation of goal-directed *behavior* instead of 'pretend' mental states.

one's *own*. They are certainly not one's own in the sense of 'differentiated from those of others.' Hutto (2007a) argues that even in their first dialogical interchanges, children have yet 'to step out of what is, in effect, a solipsistic point of view – for each child, the world is their world and any knowledge others may have of it is firmly evaluated against how they take things to stand (which, for them, is the same as how things are)' (p.210). Hutto thinks that 'solipsism' is a good label for this, and he approvingly cites Nelson (2003) who observes that 'Piaget calls this egocentrism but it is an egocentrism that simply lacks perspective because there is no possible alternative view but one's own. There are no insights into another's life because there is no vehicle except shared actions through which experience can be shared' (p.29). However, if imitation is as important to intersubjectivity as empirical studies suggest it is, then the terms 'egocentric' and 'own' take on a whole new meaning. We will further discuss this in a later chapter. Let us now take a look at what might be another precursor to simulation: *pretend play*.

Developmental findings on the ability to engage in pretend play could shed some light on the ontogeny of the capacity for simulation as well. For example, Leslie (1987) has shown that, by 2 years of age, children are already able to use a banana as if it were a telephone. The child might pick up a banana, hold it up to his ear and mouth and says: 'Hi. How are you? [Brief pause.] I'm Fine. OK. Bye.' These manifestations of pretend play are firmly rooted in very practical second-person interactions. Leslie (1994), for example, describes how child and experimenter interact in a pretend tea party. First, the child is encouraged to 'fill' two toy cups with 'juice' or 'tea' or whatever the child designates the pretend contents of the bottle to be. The experimenter then says, 'Watch this!', picks up one of the cups, turns it upside down, shakes it for a second, then replaces it alongside the other cup. The child is then asked to point at the 'full cup' and at the 'empty cup' (both cups are, of course, really empty throughout). When asked to point at the 'empty cup', 2-year-olds point to the cup that had been turned upside down.

Pretend play obviously involves not only the 'as if' condition (and some degree of inhibitory control), but also the instrumentality condition. So we might argue that it has all the ingredients to qualify as a precursor to ST. However, this by itself does not show that simulation is the cornerstone of intersubjectivity. On the contrary: if pretend play, as a precursor to simulation, develops relatively late (compared to imitation, for example), then it is reasonable to assume that the capacity for full-blown simulation is probably a quite advanced ability that develops even later. Of course, much depends on how the notion of full-blown simulation is explicated. So far I have mainly concentrated on the kind of 'high-level' ST that can be spelled out at the personal level of description. However, the problems with explicit simulation routines (such as the phenomenological objections and the problem of collapse) have lead many simulation theorists to search for a notion of 'low-level' or tacit simulation that could be fruitfully articulated at the sub-personal level.

Tacit simulation: how low can we go?

The growing attention for sub-personal processes that might support ST is in line with a more general shift in the intersubjectivity debate from high-level social understanding in terms of propositional attitudes to low-level mechanisms at the level of neurobiology. Interestingly, one of the initiators of this movement has been Goldman himself. In his 1998 paper 'Mirror neurons and the Simulation Theory of mindreading', written in collaboration with Vittorio Gallese, Goldman argued that the discovery of *mirror neurons* supported the basic tenets of his version of ST. Mirror neurons are a specific class of visuomotor neurons that fire both when one performs an action and when one observes the same action performed by another (Rizzolatti et al. 1996, 2000). The behavior of the other is 'mirrored', as though the observer himself were acting. Mirror neurons appear to be involved in a larger cortical system that matches the observation and execution of goal-related motor actions - a 'mirror neuron system'.

Initially, Gallese and Goldman (1989) conjectured that such a mirror neuron system could be seen as a 'primitive version, or possibly a precursor in phylogeny, of a simulation heuristic that might underlie mindreading' (p. 498). Mirror neuron activity seemed 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.). The mirror neuron system supported at least a kind of low-level simulation, so it was thought, but it probably also paved the way for high-level simulation in all its glory.

In more recent work, however, Gallese has distanced himself from this last idea. He now puts forward his own ST model, which is motivated by a so-called 'shared manifold' hypothesis (cf. Gallese 2001). According to this hypothesis, we are able to interact with other agents because there is a multiplicity of states that we share with them, such as emotions, body schemas and all kinds of somatic sensations. The shared manifold can be operationalized at three different levels:

(i) The *phenomenological or empathic level*, which is responsible for the sense of similarity that we experience during our meetings with other minds;
(ii) The *functional level* can be characterized in terms of simulation routines, *as if* processes enabling models of others to be created;
(iii) The *subpersonal level* is instantiated as the result of the activity of a series of mirror matching neural circuits.

According to the shared manifold hypothesis, our understanding of others is achieved by 'modeling a *behavior* as an *action* with the help of a motor equivalence between what the others do and what the observer does' (p.39, italics in original). This low-level process is automatic, unconscious and non-predicative, and Gallese (2005) argues that it *obviates* the need for complicated high-level simulation routines. 'Whenever we face situations in which exposure to others' behavior requires a response by us, be it active or simply attentive, we seldom engage ourselves in an explicit, deliberate interpretive act. Our understanding of a situation most of the time is immediate, automatic, and almost reflex like' (p.102).

Gallese is not the only one who has changed his mind. Goldman has also expressed doubts about the relevance of matching mirror neurons for a conception of simulation as being essentially a mindreading process. 'Does [Gallese's] model really fit the pattern of ST? Since the model posits unmediated resonance, it does not fit the usual examples of simulation in which pretend states are created and then operated upon by the attributor's own cognitive equipment (e.g. a decision-making mechanism), yielding an output that gets attributed to the target' (Goldman and Sripada 2005, p.207-8). Thus, the prospects for a happy marriage between the mirror neuron system and traditional articulations of high-level simulation appear to be slim. In fact, if the default way in which we understand others is indeed 'immediate, automatic, and almost reflex like', then, as Gallagher (2007) observes, this actually provides us with extra phenomenological ammunition *against* high-level ST.

However, the appeal to low-level simulation might also solve a serious problem for high-level ST. As we saw in the earlier sections, a serious problem for explicit accounts of ST is that an inference or projection of my simulation onto your mind (even with the relevant adjustments) logically still implies that I only understand *myself* in the other's situation - I don't understand *you*.[30] It is possible to extend this argument to low-level simulation: how is the mirror neuron system able to differentiate between situations in which I observe a specific goal-related behavior, and those in which I perform the same action myself? Both situations activate the same cortical sectors. Thus, a neural mechanism is needed in one of the non-overlapping brain areas to determine whether I observe or perform - whether the action is mine or yours.

A recent idea is that resonating cortical sectors or 'shared representations' are neither first- nor third-person. Our observation of goal-related behavior triggers the activation of *neutral* representations, so-called 'naked intentions' (deVignemont 2004, Jeannerod and Pacherie 2004, Gallese 2005, Hurley 2005). The mirror neuron system simulates the intention behind the action, but not the agent who executes it. The attribution of agency takes place in a second step, and is taken care of by the 'Who' mechanism (Georgieff and Jeannerod 1998). Evidence for such a neural mechanism has been found in experiments showing a differential activation in the posterior insula when the subject took the role of agent, and in the right inferior parietal cortex when it took the role of observer (Farrer et al. 2003, Farrer and Frith 2002, Ruby and Decety 2001).

These findings seem to offer low-level simulation theorists a way to circumvent the objection that the mirror neuron system is not able to differentiate between my observation of a goal-directed action and my execution of it. But Jeannerod and Pacherie (2004) make an additional, much stronger claim as well. They argue that naked intentions show up in the *phenomenology* of social interaction and can be experienced at the *personal* level: 'We can be aware of an intention, without by the same token being aware of whose intention it is' (p.140). In order to determine the author of the intention, however, we need more information. Where does this come from? 'When the naked intention one is aware of yields an overt action, the extra information needed to establish authorship may be found in the outside world. The question 'Is this intention mine?' would then be answered by answering the question: 'Is this my body performing the corresponding action?' (ibid.).

This train of thought leads to a simulation process that is structured in the following way: first, the mirror neuron system facilitates a matching process between activated cortical sectors. This results in naked intentions, which we experience at the

[30] Although Gordon argues that we have to imagine the other (and not ourselves) in his or her situation by means of a personal-level transformation, it is not really clear how this is an improvement, since I am still imagining this from *my own* point of view.

phenomenological level. Second, the ‘Who’ system determines the authorship of the action, which corresponds with an experience of authorship when it is our body that performs the action in question.

Gallagher (2007), however, has argued that the ‘who’ question hardly ever comes up at the level of experience. Most of the time, our intentions come already ‘clothed in agency’, because ‘the neural systems have already decided the issue - one way or the other - i.e., even if I'm wrong about who is acting, I am still experiencing or perceiving the intention as already determined in respect to agency’ (p.70). Moreover, Jeannerod and Pacherie seem to think that there has to be some kind of functional resemblance between the simulation processes as described at the neuronal and the phenomenological level. But this assumption of isomorphism, as I have argued in the previous chapter, is questionable (cf. chapter 1.3).

This brings us to a more severe conceptual problem for low-level simulation. It has to do with the question whether the neurobiological processes appealed to by ST in fact qualify as ‘simulation’ in the proper sense of the word. Although there are large differences between the various versions of ST, they all conceptualize simulation in a similar way. Accordingly, simulation crucially involves: (i) instrumentality, in the sense that simulation is a process I control (I *use* myself as a model), and (ii) pretense, in the sense that I put myself (‘as if’) in the shoes of the other person. Bernier (2002), for example, claims that ‘according to ST, a simulator who runs a simulation of a target would *use* the resources of her own decision making mechanism, in an ‘off-line’ mode, and then the mechanism would be fed with the mental states she would have *if* she was in the target's situation’ (p.34, italics added).

These articulations of the term simulation make sense insofar as they concern the *personal* level of description. But it is less clear whether we can explicate the notion of simulation at the *sub-personal* level without losing its original meaning. Is it meaningful to talk about pretense at the level of neurobiology? Gallese (2001) seems to answer this question affirmatively, since he argues that ‘our motor system becomes active *as if* we were executing that very same action that we are observing’ (p.37). And Gordon (2005) goes even further by saying that ‘the neurons that respond when I see your intentional action, respond "*as if*” I were carrying out the behavior […]’ (p.96). These kinds of statements are often combined with talk about instrumentality, in the sense that we are supposed to *use* our brain to model the intentional action of others. Gordon (2004), for example, claims that 'one's own behavior control system is employed as a manipulable model of other such systems. (This is not to say that the "person" who is simulating is the model; rather, only that one's brain can be manipulated to model other persons)' (p.1).

According to Gallagher and Zahavi (2008), the above attempts to attribute pretense and instrumentality to mirror neuron systems amount to *category mistakes.* They argue that it simply does not make sense to use the notion of pretense in the context of sub-personal processes. 'In sub-personal processes there is no pretense, and this is the case whether we consider neuronal processes as vehicles (mechanisms) or in terms of the content that they might represent. As vehicles, neurons either fire or do not fire. They do not pretend to fire. More to the point, however, what these neurons represent or register cannot be pretense in the way required for ST. They do not fire 'as if' *I were you*. As we saw, proponents of implicit ST claim that the mirror system is neutral with respect to the agent; there is no first- or third-person specification involved. In that case it is not possible for them to register *my* intentions as pretending to be *your* intentions' (Gallagher and Zahavi 2008, p.180; cf. Gallagher 2007, pp.360-1). The notion of instrumentality shares a similar fate. 'If simulation is characterized as a process that I (or my brain) instrumentally use(s) or control(s), if this is what simulation is, then it seems clear that what is happening in the implicit process of motor resonance is not simulation. We, at the personal level, do not *do* anything with the activated brain areas - in fact, we have no instrumental access to neuronal activation, and we can't use it as a model. Nor does it make sense to say that at the sub-personal level the brain itself is *using* a model or methodology, or *comparing* one experience with another, or *creating* pretend states, or that one set of neurons makes use of another set of neurons as a model' (ibid.).[31] As Slors (2009) has argued, the main problem here seems to be that the notion of instrumentality, despite its compatibility with the active, endogenously produced character of simulation routines, is not so easy to combine with the fact that neural resonance is often exogenously produced and has a much more passive character.

Gallagher and Zahavi think that these conceptual objections show that mirror neurons do not provide evidence for ST, period. But although their criticism might be right on target,

[31] These considerations might also shed some light on the attempt of simulation theorists to make sense of simulation at level of motor processes for action planning. It has been argued that the brain runs 'simulations' of intended movements in order to make non-conscious corrections and keep the action on track (Gallese 2001, Hurley 2005). According to Gallagher (2007), however, such a notion of simulation again fails to meet the pretense condition.

one could still maintain that mirror neurons do in fact exhibit a remarkable feature: *process replication.* Consider Goldman and Sripada (2005), for example, who have articulated a very minimal notion of simulation. They claim that we should not regard '[...] the creation of pretend states, or the deployment of cognitive equipment to process such states, as essential to the generic idea of simulation. The general idea of simulation is that the simulating process should be similar, in relevant respects, to the simulated process. Applied to mindreading, a minimally necessary condition is that the state ascribed to the target is ascribed as a result of the attributor's instantiating, undergoing, or experiencing, that very state. In the case of successful simulation, the experienced state matches that of the target' (p.208). It is clear that such a notion of simulation does not meet the conditions of pretense and instrumentality. And we might disagree about the precise extent to which our dictionary definition of simulation is applicable to resonance processes. But isn't this merely a terminological issue? Shouldn't we focus on what mirror neurons in fact *do* contribute to social interaction? Slors (2009), for example, argues that although Gallagher and Zahavi are correct in many of their observations, they cannot argue away the highly suggestive fact that resonance involves the *replication* of neural events causally responsible for intentional or emotional behavior.

However, this new claim about simulation as an instance of process replication also calls for critical review. Csibra (2005), for example, has argued that on a conservative estimation, only between 21-45% of neurons identified as mirror neurons are sensitive to multiple types of action. The motor properties of those neurons that are activated by a single type of observed action are not necessarily instantiated when the same action is actually performed. Approximately 60% of the mirror neurons are 'broadly congruent', i.e. denote a relation between an observed action and its associated executed action, but this is not an exact match. Only about 30% shows a one-to-one congruence. Newman-Norlund et al. (2007) therefore suggest that the broadly congruent mirror neurons may underlie *complementary* actions rather than *similar* actions. Although these observations do not question the importance of mirror neurons per se, they do undermine claims about simulation as a perfect match between mirror neuron processes.

There is a more important point to be made, however. It concerns the fact that the argument for process replication still takes the mirror neuron system to support a functional step-wise procedure, and assumes that it is possible to draw a strict line between the observation of an action and something that counts as a replication. Gallagher (2007) has argued, however, that if we take a closer look at the neural process involved in low-level simulation, we see that there is only a short amount of time (30-100 ms) between the activation of the visual cortex and the activation of the pre-motor cortex. And this raises the question of where exactly to draw the line between perception and replication. Perhaps even more important is what this implies: ‘Even if it is possible to draw a line between activation of the visual cortex and activation of the pre-motor cortex, this does not mean that this line distinguishes, on either a functional or phenomenological level between perception and simulation as a step-wise process [...] rather than a temporally extended and enactive perceptual process’ (p.71).

2.3 Simulation, anyone?

Before I summarize my discussion of ST so far, let me briefly comment on a popular way to frame the debate between TT and ST. It is often suggested that ST depends on a first to third-person argument, while TT depends on a third to first-person argument. Although this is not entirely untrue, we have to be cautious in associating ST too closely with the first-person perspective, and/or TT with the third-person perspective. Hurley (2005) correctly remarks that the theory versus simulation distinction cuts across acceptance or rejection of the first to third-person direction of explanation. Meltzoff’s work, for example, is often interpreted as an articulation of TT, while at the same time it also contains the analogical ‘like me’ element of ST. By contrast, Gordon’s radical version of ST explicitly rejects this analogical inference.

Simulation summarized

In this chapter I reviewed and discussed the ST approach to intersubjectivity. Since ST hails itself as the successor of TT, an important question is whether and to which extent it offers a satisfactory alternative when it comes to explaining intersubjectivity. I have shown that, insofar as ST sticks to the traditional view of intersubjectivity as crucially involving mindreading, it fails to do so and eventually collapses back into theory. Goldman’s current articulation of ST is a very clear illustration of how such a commitment naturally leads to the adoption of a hybrid model that accommodates both simulation and theorizing. But this is clearly a step back - at least insofar it amounts to a restatement of all the TT problems that initiated the whole ST movement in the first place.

Of course, ST does not necessarily have to follow the course laid down by Goldman. Heal's notion of co-cognition, for example, is much less demanding than Goldman's pretense-driven offline simulation. This is mainly because it is much more modest: it only seeks to explain how we are able to predict the thoughts of others in cases where we already posses the background knowledge required to do so. But the interesting question is precisely how we acquire this knowledge. Moreover, it is clear that Heal needs certain principles as well. These are not so much theoretical, but *rational*, and this brings along a set of new problems.

Gordon's radical simulation is probably the most promising candidate amongst the versions of ST discussed above, in particular when we include his proposal about self-attribution in terms of ascent routines. But although it seems phenomenologically sound to claim that we sometimes try to imagine ourselves in the other's shoes in order to figure out what they are thinking or feeling, it is not clear how we can explain this ability in terms of a transformation or egocentric shift at the personal level. Moreover, the fact that we sometimes use such 'Holmesian heuristics', as Hutto (2007a) calls them, does not at all imply that they are *central* to our intersubjective engagements.

There are many conceptual problems with the interpretation of the empirical evidence put forward in support of ST. If we conceptualize a notion of simulation that satisfies the pretense and instrumentality condition, then claims about high-level simulation make sense but are not supported by the evidence. We can only point at so-called 'pre-cursors', but the question is whether they suggest an interpretation in terms of simulation. As long as TT and ST are the only games in town, we might favor such a simulation interpretation over a theoretical one. But there might be other options as well. Claims about low-level simulation, on the other hand, are supported by empirical evidence but fail to make sense. There is impressive empirical evidence for the existence of resonance processes, but since mirror neurons do not satisfy the pretense or the instrumentality condition, an interpretation in terms of simulation is rather far-fetched.

Until now my discussion of TT and ST has mainly focused on certain internal problems that arise once we accept the picture of intersubjectivity they presuppose. But it is also possible to question this picture at a more basic level, in order to uncover a number of assumptions that both positions seem to have in common. This is the topic of the next chapter.

3.

Beyond the Problem of the Other Mind

The essential implications of Cartesianism for the modern self might be summed up in two words: disengagement and reflexivity.

- Sass 1992

What lurks below

In the previous chapters I have pointed out some of the internal problems with TT and ST explanations of our everyday encounters with others - problems that appear when one uncritically accepts certain assumptions about social interaction. To a large extent, these assumptions are rooted in a very influential picture of intersubjectivity that was proposed by Descartes, became problematic during the rise of British empiricism, and eventually gave birth to the problem of the other mind. The aim of this chapter is to uncover and challenge this picture of intersubjectivity.

I start by introducing the historical background of three important assumptions that have become orthodoxy for contemporary TT and ST approaches (section 1). In the first place, this is the idea that our meetings with other minds are intrinsically *problematic*, since they are deeply infused by a Cartesian phenomenology of uncertainty. Secondly, by accepting the problem of the other mind at face value, TT and ST also accept a certain conception of the mind: as a self-centered, disembodied and disembedded entity. Thirdly, they assume that our doubts about other minds can be overcome by a conscious, cognitive process - a stepwise procedure initiated by a hyper-reflexive agent.

The chapter then continues by discussing what I, following Hurley (2008), call the ‘sandwich’ model of intersubjectivity, according to which this conscious, cognitive process necessarily intervenes between our perception of the bodily behavior of other persons and our interaction with their minds (section 2). For ST, this intervention involves some version of the argument from analogy: since I know my own mind and how it relates to my body, I am able to infer that this is also true for the other on the basis of an *analogy* between our bodies. TT rejects the analogy in the argument from analogy, but it retains the inferential element. It claims that we understand others by *inferring* the contents of their minds on the basis of a theory. I will take a closer look at three components of these action-perception interventions: introspection, inference and mental concept mastery (section 3), and claim that they are problematic insofar ST and TT try to construe them as internal capacities of the individual mind.

In the final part of this chapter, I challenge what I take to be at the core of the picture of intersubjectivity presupposed by ST and TT. This is the assumption that we are normally at a theoretical remove from other people, and have to adopt a theoretical attitude towards them for the purposes of prediction, explanation and control. Instead of taking such a third-person approach as the hallmark feature of our intersubjective engagements, I propose that our meetings with other minds are primarily rooted in *second-person interactions* (section 4).

3.1 The problem of the other mind

The Cartesian picture of the mind

In order to get a clearer view of the problems troubling TT and ST, we have to address the deeper assumptions that they have in common. Gallagher (2004) argues that both positions share two important presuppositions: the ’mentalistic supposition’ and the ‘supposition of universality’. I take the mentalistic supposition as the starting point for my diagnosis:

*Supposition 1 (the mentalistic supposition): ‘*The problem of intersubjectivity is precisely the problem of other minds. That is, the problem is to explain how we can access the minds of others. This is a problem of access because other minds are hidden away, closed in, behind the overt behavior that we can see. This is a mentalistic and clearly Cartesian supposition about the very nature of what we call the mind. The mind is conceived as an inner realm, in contrast to behavior, which is external and observable, and which borrows its intentionality from the mental states that control it. Both TT and ST set the problem as one of gaining access to other minds, and their explanations of social cognition are framed in precisely these terms' (p.200).

To start with, we have to restrict the scope of Gallagher's claim. It is true that those TT, ST and hybrid TT/ST positions that explain intersubjectivity in terms of *mindreading* (understood as the structural attribution of mental states such as beliefs and desires) are committed to the mentalistic supposition. However, as we saw in the previous chapter, there are also lightweight versions of ST that discard the traditional ingredients of mindreading (e.g., Gordon) and/or stress the embodied nature of our understanding of others (e.g., Gallese). Since these ST approaches often explicitly reject the mentalistic supposition and many of the other assumptions that will be discussed below, they are not the target of my criticism. At the same time, however, it is not always clear whether these positions are best interpreted in terms of *simulation*. This is why I will not group them under the general header of ST. Instead, I use this label primarily to refer to the classic simulation approaches that revolve around a robust notion of mindreading.

Returning to the topic at hand, Gallagher is right that the mentalistic supposition is a Cartesian supposition. But this is certainly not the whole story. For Descartes was not yet troubled by worries about the minds of his fellow human beings. It was only against the background of British empiricism that the problem of the other mind was recognized as an 'official' philosophical problem. The genealogy of this problem warrants more detailed investigation, since it might give us a clue as to where we should look for a solution. Let us therefore briefly consider the Cartesian picture of the mind.

Descartes is well-known for his quest for certainty. What is remarkable about this quest is that it begins with a method of radical *doubt*. Descartes writes that this method imitates that of the architect. 'When an architect wants to build a house which is stable on ground where there is a sandy topsoil over underlying rock, or clay, or some other firm base, he begins by digging out a set of trenches from which he removes the sand, and anything resting on or mixed in with the sand, so that he can lay his foundations on firm soil. In the same way, I began by taking everything that was doubtful and throwing it out, like sand' (Replies 7, AT VII 537).[32]

The Cartesian method of doubt requires a highly reflexive attitude of *disengagement*, since we can only avoid mistakes and achieve certainty if we suspend our judgment and 'hold back', assenting only to that which we can clearly and distinctly perceive to be true. This leads Descartes to the conclusion that sky and earth, colours and sounds, and in fact all external things are nothing better than the illusions of dreams, and he even comes to consider himself as without hands, eyes, or any of the senses, and as falsely believing that he is in possession of these. Eventually, however, he manages to find something that lies beyond all doubt. This is the famous 'cogito ergo sum'.[33] It is important to notice that the cogito (the 'I think') is not the result of an *inference*.[34] Instead, it is recognized by an inner awareness - a simple and immediate act of clear and distinct perception. The cogito is the unifier of all modes of thinking (doubting, dreaming, understanding, willing etc.) and it provides Descartes with a foundation upon which to build further: 'Archimedes used to demand just one firm and immovable point in order to shift the entire earth; so I too can hope for great things if I manage to find just one thing, however slight, that is certain and unshakable' (AT VII 24, CSM II 16). This, however, comes at a steep price. For the cogito is identified as an immaterial and timeless substance, and radically cut off from body and world. It becomes a passive spectator, separated from its natural and social context and no longer situated in culture or language.

[32] References to Descartes' work are abbreviated as follows: AT: *Oeuvres de Descartes*. 1904. Adam C. and Tannery T. (eds.) Paris: Vrin; CSM: *The Philosophical Writings of Descartes Volumes I and II*. 1984. Cottingham J., Stoothoff R. and Murdoch D. (eds.) Cambridge: Cambridge University Press; CSMK: *The Philosophical Writings of Descartes Volume III.* 1984. Cottingham J., Stoothoff R., Murdoch D. and Kenny A. (eds.) Cambridge: Cambridge University Press. In citations of AT, CSM, and CSMK, Roman numerals refer to volume and Arabic numerals to page.

[33] Descartes writes that 'I have convinced myself that there is absolutely nothing in the world, no sky, no earth, no minds, no bodies. Does it now follow that I too do not exist? No: if I convinced myself of something then I certainly existed. But there is a deceiver of supreme power and cunning who is deliberately and constantly deceiving me. In that case I too undoubtedly exist, if he is deceiving me; and let him deceive me as much as he can, he will never bring it about that I am nothing so long as I think that I am something. So after considering everything very thoroughly, I must finally conclude that this proposition, *I am*, *I exist*, is necessarily true whenever it is put forward by me or conceived in my mind' (Med. 2, AT VII 25).

[34] Descartes remarks that: 'When someone says "I am thinking, therefore I am, or I exist," he does not deduce existence from thought by means of a syllogism, but recognizes it as something self-evident by a simple intuition of the mind' (Replies 2, AT VII 140).

The Cartesian conception of the mind as self-founded and locked into itself has become established as the official doctrine of the modern mind. This idea, which Damasio calls 'Descartes' error', fuels the mentalistic supposition and motivates contemporary notions of the mind as a disembodied entity, hidden away and closed in behind overt behavior. It also raises the question of how we can have *access* to such a mind.

On the Cartesian view, this works as follows. When it comes to my own mind, I have a kind of so-called 'privileged access': an immediate and intuitive awareness of my inner life. Although I may start out being in a state of confusion or error, I have the ability to turn inwards and perceive the contents of my mind with utter clarity, reflecting in a methodological manner upon my stream of consciousness. Such a clear and distinct introspection, guided by the 'great light in the intellect', is illuminating and provides me with intimate knowledge of the mind's ideas. These ideas are innate and universal - they represent 'true, immutable and eternal essences' (CSMK 183, AT III 383), and Descartes writes that they have 'a seat in our mind' (CSMK 23, AT I 145). My access to the minds of others, however, is always mediated by their bodily behavior. And my perception of this behavior, like sense perception in general, is potentially *misleading*. Descartes observes that we 'misuse them [the senses] by treating them as reliable touchstones for immediate judgements about the essential nature of the bodies located outside us; yet this is an area where they provide only very obscure information' (CSM II 57-58, AT VII 83). What we perceive through our external senses results at best in a 'spontaneous impulse' to believe something.

This, of course, presses the question how we are able to access *other* minds. Since these are not 'presented' to me in the way my own mind is, they have to be 'represented'. This, however, is not really a problem for Descartes. We have privileged access to the ideas in our own mind, and it is through our knowledge of these ideas that we are in touch with the minds of our fellow human beings. Self-knowledge provides a secure basis for our knowledge of others. Importantly, we do not have to infer the existence of their minds on the basis of an analogy. Instead, Descartes short-circuits the problem of the other mind with an argument from faith. He claims that God has created man in such a way that our ideas truthfully represent what is out there in the external world, including the other's mind.

The above picture of the mind has been decisive for contemporary views on intersubjectivity, and we can find many Cartesian elements in both TT and ST approaches. One of them is the notion of a disembodied mind - hidden away and closed in behind the overt behavior that we can see. But this notion is inevitably the result of an attitude of *disengagement* that eventually puts one at a distance from practice, where the usual clues on which we rely to orient ourselves and make sense of things are no longer available. And this attitude, in turn, has to be seen in the context of a *phenomenology of uncertainty*, in which we constantly doubt everything that occurs around us – including the intentions and behaviors of others. In such a context, it is very tempting to propose that the resulting gap between doubt and certainty has to be bridged by a *theoretical intervention.*

The phenomenology of uncertainty, our disengaged stance towards others, and the theoretical attitude by which we are to overcome our doubts are all part of the picture of intersubjectivity that is presupposed by both TT and ST. So is the idea that this attitude is universally acquired by all human beings. This supposition bears similarities to the Cartesian postulate of innate ideas that have a universal status. Gallagher (2004) calls it the supposition of universality:

Supposition 2 (the supposition of universality): 'Our reliance on theory (or our reliance on simulation or some combination of theory and simulation) is close to universal. That is, this folk-psychological way of understanding and interacting with others is pervasive in our everyday life' (p.200).

However, there is still another aspect of Cartesianism that has been very influential in shaping the intersubjectivity debate. According to Descartes, certain knowledge requires that we clearly and distinctly perceive with the mind's eye. He claims that 'doubtless, there is nothing that gives me assurance of [...] truth except the clear and distinct perception of what I affirm, which would not indeed be sufficient to give me the assurance that what I say is true, if it could ever happen that anything I thus clearly and distinctly perceived should prove false; and accordingly it seems to me that I may now take as a general rule, that all that is very clearly and distinctly apprehended (conceived) is true' (Med. 3, AT VII 35).[35] On the Cartesian view, true knowledge is modelled on a clear and distinct perception of the individual mind.

This results in what Dewey (1960, p.23) calls a 'spectator' theory of knowledge: 'the theory of knowing is modeled after what was supposed to take place in the act of vision.

[35] Somewhere else Descartes writes that 'My nature is such that so long as I perceive something very clearly and distinctly I cannot but believe it to be true' (Med. 5, AT VII 69).

The object refracts light and is seen; it makes a difference to the eye and to the person having an optical apparatus, but none to the thing seen. The real object is the object so fixed in its regal aloofness that it is a king to any beholding mind that may gaze upon it. A spectator theory of knowledge is the inevitable outcome.' According to Descartes, to know is to clearly and distinctly perceive the immediate and the intuitive (the ideas), and to suppress spontaneous impulses to believe something merely on the basis of external sense perception. 'Real' perception, in the Cartesian sense, is not a kind of action. It merely aims to reflect ideas without altering them. It is the *passive* recognition of something that is already there. To perceive with the mind's eye is to remain still and impartial – not actively engaged in the process of perceiving.

Dewey argues that a spectator theory of knowledge inevitably leads to a strict separation between perception, thinking and action. In the case of Descartes, it results in a distinction between (i) sensory perceptions (or bodily sensations), (ii) 'thinking', or a clear and distinct perception of mental ideas, and (iii) behavioral responses. In his discussion of the concept of the reflex arc in psychology, Dewey (1896) complains that these old distinctions are still firmly in place: 'instead of interpreting the character of sensation, idea and action from their place and function in the sensory-motor circuit, we still incline to interpret the latter from our preconceived and preformulated ideas of rigid distinctions between sensations, thoughts and acts. The sensory stimulus is one thing, the central activity, standing for [representing] the idea, and the motor discharge, standing for [representing] the act proper, is a third. As a result, the reflex arc is not a comprehensive, or organic unity, but a patchwork of disjointed parts, a mechanical conjunction of unallied processes' (p.358).

In contemporary discussions about intersubjectivity, the boundaries between perception, thinking and action are often still in place as well. And although there are signs that they are slowly dissolving, many proponents of TT and ST still maintain that our understanding of others somehow has to follows a Cartesian perception-thinking-action route. Of course, they have different ideas as to how the specific steps of this route should be explicated. This brings us to the next stage in the development of the problem of the other mind.

3.2 Empiricism and the argument from analogy

Although Descartes argues that clear and distinct perception is by far the best candidate for knowledge, he realizes it still falls short of absolute certainty. Descartes points out that, in order to achieve absolute certainty, he has to overcome the Evil Genius doubt and prove the existence of a non-deceiving God. ‘But, that I may be able wholly to remove it, I must inquire whether there is a God, as soon as an opportunity of doing so shall present itself; and if I find that there is a God, I must examine likewise whether he can be a deceiver; for, without the knowledge of these two truths, I do not see that I can ever be certain of anything’ (Med. 3, AT VII 36). Unsurprisingly, Descartes eventually comes to the conclusion that there is a God and that He is no deceiver. This enables him to evade the solipsistic consequences of his method of doubt and neutralize the problem of the other mind.

The problem of the other mind did not come to the fore until the rise of British empiricism, when the appeal to a benevolent God was no longer taken for granted and a number of other Cartesian commitments became unacceptable as well. Locke, for example, still accepted the essentials of the Cartesian picture of the mind, but rejected the claim that some truths must be innate because they are universally understood. He pointed out that the universality of a certain truth does not imply that it is therefore necessarily innate, for it could have been *learned* by all people. Moreover, the fact that infants and the mentally impaired do not understand them testifies against the plausibility of universal innate ideas. Contrary to Descartes, Locke believed that the mind of a person at birth is a tabula rasa, a blank slate upon which knowledge is imprinted through experience. He argued that ideas are derived from experience either by sensation (the affection of the senses through the observation of external bodies) or reflection (the perception of the operations of our own mind). By rejecting innateness, Locke had all the ingredients to conjure up the problem of the other mind. However, he seems not to have recognized this.

It is generally thought that, as such, the problem of the other mind was not recognized until John Stuart Mill (1878) explicitly articulated it as a prominent philosophical issue by asking: ‘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 figures which I see and hear, have sensations and thoughts, or in other words, possess Minds?’ (p.243). This is not correct, however, since it was actually Thomas Reid who was the first to identify the problem of the other mind (cf. Avramides 2001).[36] It is also generally accepted, correctly this time, that Mill was the first to propose the infamous argument from analogy as a solution to this problem.[37]

Mill argued that, by observing that the bodies of other human beings behave as my body does in similar circumstances, I am able to infer that the mind I know to accompany my bodily behavior is also present in the case of others. '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 acts, and other outward signs, which in my own case I know by experience to be caused by feelings. I am conscious in myself of a series of facts connected by a uniform sequence, of which the beginning is modifications of my body, the middle is feelings, the end is outward demeanor. In the case of other human beings I have the evidence of my senses for the first and last links of the series, but not for the intermediate link. I find, however, that the sequence between the first and last is as regular and constant in those other cases as it is in mine […] I must either believe them to be alive, or to be automatons: and by believing them to be alive, that is, by supposing the link to be of the same nature as in the case of which I have experience, and which is in all other respects similar, I bring other human beings, as phenomena, under the same generalizations which I know by experience to be the true theory of my own existence' (1878, p.243).

The argument from analogy crucially depends on the ability to make inferences such as 'if there is a modification of my body of kind B, then usually an experience of kind E is occurring as well', or 'if there is an experience in my mind of kind E, then usually this causes a bodily reaction of kind R.' In our own case, according to Mill, these psycho-behavioral generalizations are available because we are 'conscious' of the proper connections between (a) the modifications of my body, (b) my feelings, and (c) my outward

[36] Indeed, it seems that the first frequent use of the words 'other minds' is to be credited to him (Somerville 1989, p.249).

[37] The idea that we understand others by means of *inference* was already introduced by David Hume, who wrote that 'no passion of another discovers itself immediately to mind. We are only sensible of its causes or effects. From these we infer the passion: And consequently these give rise to our sympathy' (2003, p.410). But this answer only postpones the difficulty: by what *sort* of inference do we understand other minds? Hume is not of much help here, though it is clear that he thinks of the reasoning in terms of causes and effects: whatever inferences they are, they are based on laws or regularities which we have learned through experience hold in experience.

demeanor. In case of other minds (a) and (c) are present, but (b) is missing. However, if the connection between (a) and (c) is of the same nature as in my own case, then *by analogy* we have reason to expect them to be just as regular and constant.

Although the argument from analogy still retained the Cartesian appeal to introspection and the Cartesian primacy of self-knowledge, British empiricism rejected not only the existence of innate ideas but also abandoned the Cartesian search for absolute certainty. This, however, led to the following question: how sure can we actually be of the existence of the other mind? Consider Bertrand Russell’s formulation of the argument from analogy, for example. Russell (1948) initially proposed that ‘from subjective observation I know that A, which is a thought or feeling, causes B, which is a bodily act, e.g., a statement. I know also that, whenever B is an act of my own body, A is its cause. I now observe an act of the kind B in a body not my own, and I am having no thought or feeling of the kind A. But I still believe, on the basis of self-observation, that only A can cause B; I therefore infer that there was an A which caused B, though it was not an A that I could observe. On this ground I infer that other people's bodies are associated with minds, which resemble mine in proportion as their bodily behavior resembles my own’ (p.486).

However, Russell soon realized that the argument from analogy, thus formulated, is only applicable in *idealized* circumstances. In practice, ‘the exactness and certainty of the above statement must be softened’, because even in our own case we cannot be sure that A is the only cause of B. It is possible that, although we experience A to be the cause of B, there are other causes of B ‘outside our experience’. Therefore, Russell also offered a ‘common sense’ version of the argument of analogy: ‘If, whenever we can observe whether A and B are present or absent, we find that every case of B has an A as a causal antecedent, then it is *probable* that most B's have A's as causal antecedents, even in cases where observation does not enable us to know whether A is present or not’ (ibid., italics added). In other words, Russell weakened the conclusion of the argument from analogy because he doubted the human capacity for self-observation. But he still thought the argument itself was basically correct as a solution to the problem of the other mind.

The same is true for Theodor Lipps, who also remained loyal to the argument from analogy. According to Lipps, however, our understanding of others is not based on a conscious, inferential process that begins with the clear perception of our own mind. Instead, it depends on an unconscious process of *empathy*, in which we project ourselves into the physical manifestations evinced by others.[38] Lipps (1993) suggested that such a process involves an element of 'inner imitation', and it is driven by our 'natural instinct' When watching an acrobat on a tightrope, for example, the perceived movements and affective expressions of the acrobat are ‘instinctively’ and simultaneously mirrored by kinesthetic ‘strivings’ and experiences of corresponding feelings in the observer.

Lipps used the notion of empathy in order to stress the affective, bodily and experiential dimension of how we understand others. Although empathic understanding is still based on an analogy, it does not necessarily require that we are always *aware* of how our own mind relates to our body, or that we are continuously busy *inferring* that the same is also true for other persons. In this respect, it clearly gives us a more parsimonious phenomenological account of everyday intersubjectivity.

However, the solutions offered by Russell and Lipps do not really show how the argument from analogy is able to provide us with *reliable* knowledge of the other mind. Although many philosophers no longer care about absolute knowledge Cartesian-style, they do find it problematic that the argument is based on an inductive generalization from only *one* case. Paul Churchland (1988), for example, has argued that this makes it the weakest possible instance of an inductive argument.[39] He thinks it is possible to overcome

[38] Lipps’ ideas about empathy resulted from his translation of David Hume's ‘A Treatise of Human Nature' into German, although Hume actually used the term ‘sympathy’ to describe what Lipps was interested in. Hume suggested that ‘the minds of men are mirrors to one another, not only because they reflect each other’s emotions, but also because those rays of passions, sentiments and opinions may be often reverberated, and may decay away by insensible degrees’ (2003, p.259). He argued that this was made possible by the sole principle of all passions: sympathy. ‘No quality of human nature is more remarkable, both in itself and in its consequences, than that propensity we have to sympathize with others, and to receive by communication their inclinations and sentiments, however different from, or even contrary to our own’ (p.225). Sympathy makes it possible to ‘enter deeply into the sentiments of others’, and their affections are ‘rendered present to us by the imagination’, operating as if originally our own. ‘We rejoice in their pleasures, and grieve for their sorrows, merely from the force of sympathy’ (p.277).

[39] Some philosophers have tried to avoid this objection by arguing that the argument from analogy should be based on the multitude of correlations between mental states and behavior that one observes in one's own case, rather than on a generalization proceeding from just one observed case. Ayer (1956) for example, suggests that ‘The objection that one is generalizing from a single instance can perhaps be countered by maintaining that it is not a matter of extending to all other persons a conclusion which has been found to hold for only one, but rather of proceeding from the fact that certain properties have been found to be conjoined in various circumstances. So the question that I put is not: Am I justified in assuming that what I have found to be true only of myself is also true of others? but: Having found that in various circumstances the possession of certain properties is united with the possession of a certain feeling, does this union continue to

this problem by adopting a different standard of theoretical *justification*. Churchland points out that the problem of the other mind was first formulated at a time when our grasp of the nature of theoretical justification was still rather ‘primitive’. It was believed that a general law could be justified only by an inductive generalization from a suitable number of observed instances of the elements comprehended by this law. But this only works for *observable* things and properties, while modern science is full of laws that govern the behavior of *unobservable* things and properties. These laws require a different form of empirical justification. Churchland notices that contemporary theorists postulate unobservable entities and specific laws governing them, because occasionally this produces a theory that allows them to construct predictions and explanations of observable phenomena hitherto unexplained. More specifically, they assume certain hypotheses and conjoin with them information about observable circumstances in order to deduce statements about further observable phenomena, statements which are systematically true. This is commonly called ‘hypothetico-deductive’ justification.

Churchland claims that it is precisely this kind of justification that allows us to solve the problem of the other mind. The idea is that we understand others by employing a folk psychological theory - a network of general laws connecting mental states with perceptions, bodily behavior and other mental states. These laws are plausible for the same reason that the laws of any theory are plausible: their explanatory and predictive power. The existence of the other mind is a *hypothesis*, which is plausible to the extent that the other’s behavior can be explained and predicted in terms of desires, beliefs, perceptions, emotions and so on. If this is the best way to understand the behavior of most humans, then one is justified in believing that they are ‘other minds’.

Churchland can be seen as an early adaptor of the TT approach to folk psychology. He argued that folk psychology is successful as a theory if it allows us to ‘explain and predict the behavior of human beings better than any other hypothesis currently available’

obtain when the circumstances are still further varied. The basis of the argument is broadened by absorbing the difference of persons into the difference of the situation in which the psycho-physical connections are supposed to hold’ (p.249). However, this counterargument does not work. Despite that we now have a multitude of correlations, the simple fact remains that not all instances of behavior we observe in our own case are accompanied by mental states. So the conclusion to be drawn, were we proceeding from this multitude of correlations, could only be that many instances of behavior are associated with mental states. But this is not the conclusion we need. For such a conclusion is still compatible with the idea that some of the human bodies we encounter behave just as our own body does, without being associated with mental states and thus without having a mind (cf. Hyslop and Jackson 1972).

(1988, p.71). Importantly, this does not require the examination of our own case. It is the success of our folk psychology with respect to the behavior of people *in general* that matters. Nor does it require an element of analogy, in the sense that the other is 'like me'. In fact, the other might be quite different. But this, Churchland argues, does not affect my 'theoretical access' to their 'internal states', since one could 'simply use a different psychological theory to understand their behavior, a theory different from the one that comprehends one's own inner life and outer behavior' (ibid.).[40]

Churchland frames the problem of the other mind as an *inference to the best explanation*: an inference which is guided by a folk psychological theory, bringing us from observed behavior to a hidden mental state. Although this inference does not provide us with certain knowledge of the other mind, at least it gives me more reason to believe in its existence than to deny it. But the question is whether a folk psychological theory gives us the best explanation (cf. chapter 1.5). Churchland thinks this is not the case, and he dismisses folk psychology as an empirically and conceptually degenerating research program that needs to be terminated in favor of its superior alternative: cognitive neuroscience. Other proponents of TT usually do not go as far as Churchland, and instead adjust their standards of justification. They frame the problem of the other mind in terms of *adequacy*. Although we are certainly not infallible, it is very often the case that folk psychology allows us to successfully predict what others are going to do, or explain what they have done.

3.3 Deconstructing the argument from analogy

So far I have sketched (a part of) the historical background of the problem of the other mind and the argument from analogy. We saw that the problem of the other mind encompasses more than just a notion of the mind as a disembodied and disembedded entity. At a far more profound level, it is inspired by a Cartesian anxiety, and a longing for certainty that has to be met by methodological thinking.

[40] Notice that Churchland's solution is different from the one offered by other proponents of TT, who argue that we employ the same folk psychological theory in case of self and other knowledge (cf. chapter 2).

This anxiety is still present in contemporary TT and ST explanations of intersubjectivity. It suggests that our encounters with our fellow human beings are essentially *problematic*, since we are always in the dark about their intentions, feelings and beliefs. In order to overcome our doubt in these situations, we need to take a step back and disengage from active participation. We need to adopt a theoretical, third-person stance towards others in order to figure out what they are up to, ascribing causally efficacious inner mental states to them for the purpose of prediction, explanation and control.[41] As a result, we are not actively involved but rather stand as passive observers at the margins of the situation. We do not have the slightest clue about what is going on, or how we need to *respond* to what happens, unless we call forth a theory or run a simulation routine. Proponents of both ST and TT think that we need some kind of *intervention* between our initial observation of others and our final reaction towards them.

This separation between perception and action can be seen as a consequence of the Cartesian spectator theory of knowledge, and it leads to a 'sandwich model' of intersubjectivity. Hurley (2008) argues that such a model 'regards perception as input from the world to the mind, action as output from the mind to the world, and cognition as sandwiched in between. Central cognition, on this view, is where all the conceptually structured general purpose thinking happens: perceptual information is assessed in light of standing beliefs and goals, deliberative and inferential processing occurs, action plans are formulated and sent on for execution' (p.2). According to ST, this cognitive intervention proceeds according to some version of the argument from analogy: since I know my own mind and how it relates to my body, I am able to infer that this is also true for the other on the basis of an *analogy* between our bodies. TT, by contrast, rejects the analogical element but sticks to the idea of an intervention based on *theoretical inference*. It claims that we understand others by inferring the contents of their minds on the basis of a folk psychological theory. In what follows, I will deconstruct the argument from analogy into three components: introspection, inference and mental concept mastery, and argue that these components are problematic insofar they come with serious developmental constraints and are modeled on the minds of *individual* agents.

[41] This is what Bogdan (1997) labels 'the spectatorial view of interpretation', since it portrays 'the subject as a remote object of observation and prediction' (p.104).

Introspection

In the previous chapter we already encountered a number of initial objections to the argument from analogy by philosophers such as Ryle and Scheler (cf. chapter 2.3). It pays to follow Scheler a bit further here, since he not only provides us with a whole list of direct criticisms of the argument from analogy, but also attempts to dismantle two crucial presuppositions behind it.

First, the argument from analogy assumes that we perceive only the bodies of others and therefore have to *infer* the existence of their minds. As a result, we are unable to experience the thoughts, feelings and emotions of others in a direct way. According to Scheler, however, this assumption is not supported by the phenomenological evidence. On the contrary, it is a 'phenomenological fact' that we perceive other minds, much like we perceive our own mind. Rather than being busy with inferring their mental states, we are able to directly perceive them. Scheler (1973) famously claims that '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. If anyone tells me that this is not 'perception' [...] I would beg him to turn aside from such questionable theories and address himself to the phenomenological facts' (p.254). This argument is directed against the traditional idea that perception and action require the intervention of cognition.

Second, the argument from analogy is grounded in the assumption that self-knowledge is 'given' to us in our first-person experience and can be used as a foundation for our knowledge of others. This is doubtful as well, according to Scheler, for 'who can say that it is our own individual self and its experiences which are "immediately given" in that mode of intuition, by which alone the mental, a self and its experiences, can possibly be apprehended, namely in inner intuition or perception? Where is the phenomenological evidence for this assertion?' (p.244). Scheler suggests that the argument from analogy 'underestimates the difficulties involved in self-experience and overestimates the difficulties involved in the experience of others' (ibid.).[42]

[42] Scheler's objection is similar to Sellars criticism of the myth of the given (cf. chapter 1.2). However, where Sellars and his TT followers maintained that self and other knowledge are equally

The idea that we need to introspect an inner mental realm before we can engage in social interaction is problematic when we consider our everyday phenomenology, as I remarked in the previous chapter. But according to Scheler, there is another problem as well. This has to do with the *unreliability* of introspection. The fact that for a long time this has been overlooked is partly due to the strong influence of the Cartesian ideal of introspection as a clear and distinct perception. The founders of psychology, Wilhelm Wundt and William James, were still convinced that introspection was of crucial importance for our knowledge of the mind. James, for example, said that 'the word introspection need hardly be defined – it means, of course, the looking into our own minds and reporting what we there discover. Everyone agrees that we there discover states of consciousness' (James 1890/1981, p.85). Back then, it was still thought that introspection, as a method, distinguished psychology from the natural sciences. Hempel (1949) describes the received view at the time as follows: 'It is impossible to deal adequately with the subject matter of psychology by means of physical methods. The subject matter of physics includes such concepts as mass, wave length, temperature, field intensity, etc. In dealing with these, physics employs its distinctive method which makes a combined use of description and causal explanation. Psychology, on the other hand, has for its subject matter notions which are, in a broad sense, mental. They are *toto genere* different from the concepts of physics, and the appropriate method for dealing with them scientifically is that of empathetic insight, called 'introspection', a method which is peculiar to psychology' (p.375).

However, with the rise of behaviorism, psychologists became increasingly doubtful about the prospect of introspection as a viable psychological method. Watson (1913), for example, published a statement of behaviorist principles that began as follows: 'psychology as the behaviorists view it is a purely objective experimental branch of natural science. Its theoretical goal is the prediction and control of behavior. Introspection forms no essential part of its methods, nor is the scientific value of its data dependent upon the readiness with which they lend themselves to interpretation in terms of consciousness' (p.158).

A more recent and very influential critique of introspection can be found in an article by Nisbett and Wilson (1977), who concluded that people have little or no introspective access to higher order cognitive processes. The authors reported evidence of subjects

problematic and in need of *theory,* Scheler argues that the *practice* of self and other experience is well established.

confabulating stories about the cause of the mental states they were entertaining. At a shopping mall, they mounted a display table with four pairs of identical pantyhose, labeled A, B, C and D from left to right, and asked passersby which pair they preferred and what reasons they had for doing so. In a previous version of the study, they had ascertained that there was a strong position effect: pair A was preferred by 12 percent of the participants, pair B by 17 percent, pair C by 31 percent and pair D by 40 percent. In the main study, when people where asked the reason for their choice, people pointed to some attribute of the preferred pair, such as its superior knit, sheerness, or elasticity. Nobody spontaneously mentioned the position effect as the cause of his preference – even when specifically asked whether their choice had been influenced by position (with the exception of a participant who was taking psychology courses). The authors concluded that participants seemed totally unaware of what was in fact the cause of their preference, and their claim about what caused it was merely a *confabulation*.

This is only the tip of the iceberg, and there have been many more studies on confabulation since. Gazzaniga (1992) and Bayes and Gazzaniga (2000), for example, have provided evidence for confabulation in split-brain patients, who had undergone surgical separation of their two hemispheres.[43] And Wegner (2002) has argued that confabulation, or what he calls 'intention invention', is also pervasive when it comes to our everyday self-ascription of consciously willed decisions. Whenever we explain our acts as the outcome of our conscious choice, we engage in intention invention, because our actions actually stem from countless causes of which we are completely unaware.[44] Wegner claims that 'When we apply mental explanations to our own behavior-causation mechanisms, we fall prey to the impression that our conscious will causes our actions. The fact is, we find it enormously seductive to think of ourselves as having minds, and so we are drawn into an intuitive appreciation of our own conscious will [...] The real causal sequence underlying human behavior involves a massively complicated set of

[43] These studies began as investigations of the abilities of people who have had their left and right brains surgically severed as a treatment for severe seizures. Such a treatment leaves mid and lower brain structures joining the two sides intact, but it creates a 'split brain' at the cortex.

[44] For example, consider the following study by Brasil-Neto et al. (1992). The experimenters exposed the participants to TMS (transcranial magnetic stimulation) of the motor area of the brain as the participants chose freely whether to move their right or left index finger. Surprisingly, although the participants showed a marked preference to move the finger contra-lateral to the site stimulated, they continued to perceive that they were voluntarily choosing which finger to move.

mechanisms' (2002, p.26f). This implies, according to Wegner, that an agent cannot be the real cause of his or her action. The agent self is only a virtual entity, an 'apparent mental causer' (2005, p.23).[45]

I certainly do not wish to defend Wegner's explanation of our everyday explanation of our own behavior, but I do think that the above experiments at the very least indicate that the commonsense use of introspection is far removed from the Cartesian ideal of clear and distinct perception. At the same time, however, the studies mentioned above do not seem to prove that we have no privileged access *whatsoever*. Although it might be true that we are not aware of the causes of our behavior, it could still be argued we do have a kind of privileged access to a great deal of information about ourselves, such as the content of our current thoughts and feelings, and the objects of our attention. Wilson (2002) has recently recanted part of his earlier confabulation story by admitting that 'the fact that people make errors about the causes of their responses does not mean that their inner worlds are a black box. I can bring to mind a great deal of information that is inaccessible to anyone but me. Unless you can read my mind, there is no way you could know that a specific memory just came to mind, namely an incident in high school in which I dropped my bag lunch out a third-floor window, narrowly missing a gym teacher who happened to walk around a corner at just the wrong time. Isn't this a case of having privileged "introspective access to higher order cognitive processes?" [...] Although we often have access to the results of these processes- such as my memory of the lunch-dropping accident- we do not have access to the mental processes that produced them. I don't really know, for example, why that particular memory came to mind, just as the participants in the panty-hose study did not know exactly why they preferred pair D over A' (p.150). And in a further passage, he claims that: 'To the extent that people's responses are caused by the adaptive unconscious, they do not have privileged access to the causes and must infer them, just as Nisbett and I argued. But to the extent that people's responses are caused by the conscious self, they have privileged access to the actual causes of these responses; in short, the Nisbett and Wilson argument was wrong about such cases' (p.106).

I think Wilson is correct in claiming that we do have some kind of privileged access to the contents of our own mind. But the last passage above is confusing, because it

[45] See also Wegner (2003), where he states that 'The theory of apparent mental causation turns the everyday notion of intention on its head [...] The theory says that people perceive that they are intending and that they understand behavior as intended or unintended - but they do not really intend' (p.10).

suggests that some responses are caused by the adaptive unconscious while others are caused by the conscious self. Later in this book, I will offer an alternative story about self-knowledge as an active and constructive process of *interpretation* (that also involves a certain amount of confabulation) instead of a passive introspection of one's own mental states (cf. chapter 5.2).

It is often claimed that the unreliability of introspective (phenomenological) properties poses a potential problem for those simulation theorists who rely on introspection to get their simulation routines off the ground. But a far more serious developmental constraint on the appeal to introspection is the fact that it presupposes *mental concept mastery*. If the introspection of our own mental states is the starting point for our intersubjective engagements, then this already presupposes that we are able to *identify* and *self-attribute* them. And if we are to distinguish between and clearly recognize the many varieties of mental states, thereafter to divine the connections they bear to our behavior, we must possess the concepts necessary for making such identifying judgments. We must grasp the meaning of the terms 'belief', 'desire', 'pain' and so forth. As we saw in the previous chapter, Goldman (2006) tries to avoid this requirement by putting forward *neural* states as suitable candidates for introspection. But it does not seem to make much sense to claim that we are able to introspect neural states in a conscious manner. Nor does it make sense to talk about the *unconscious* introspection of neural states, unless this process is construed as a kind of feedback or forward comparator (cf. chapter 4.3). In this case, however, it is not clear why the label 'introspection' should be used. In other words, if we insist on appealing to introspection, it seems we need a story about the acquisition of mental concepts in ontogeny.

Inference

Such a story about mental concept acquisition is also required if we wish to properly explain how human agents are able to make *inferences.* The latter ability is crucial for TT explanations of everyday social interactions, according to which we make sense of each other's actions by means of a folk psychological theory that specifies how beliefs and desires combine to give rise to intentions and actions. TT argues that this theoretical 'system of inferences' is the engine of everyday interpersonal understanding - even though it must be supported by further auxiliary generalizations about what people typically do in a range of circumstances. At the core of the theory is the belief-desire principle: ‘if A wants p and believes that doing q will bring about p, then ceteris paribus, A will q’ (Borg 2007, p.6).[46] If we support this principle with other folk psychological generalizations such as ‘persons who *want* to quench their thirst and *believe* that drinking water will satisfy their thirst, will tend to drink water’, we can construct inferential arguments and use their conclusions for the purposes of behavior *prediction*:

1. Persons who want to quench their thirst, and believe that drinking water will satisfy their thirst, will tend to drink water (folk psychological law)
2. This person feels thirsty (first premise)
3. This person believes that drinking water will satisfy his thirst (second premise)
4. Normal conditions obtain (ceteris paribus)

———————

5. This person is going to drink water (conclusion)

But we can also use these folk psychological generalizations to *explain* behavior. The question ‘Why is he drinking water?’ can be answered by referring to a belief-desire pair: ‘Because he *wants* to satisfy his thirst, and he *believes* that drinking water will satisfy his thirst’. In both cases, we infer the conclusion from a folk psychological law, in combination with the starting premises (the initial conditions needed to connect this law to the specific explanation or prediction) and the ceteris paribus clause. It is often suggested that additional principles are needed in order to guarantee that we make these inferences in a *reliable* way. Botterill (1996), for example, gives us the following principle: ’[Inference Principle] When an agent A acquires the belief that p and a rational thinker ought to infer q from the conjunction of p with other beliefs that A has, A comes to believe that q’ (p.116).

As I already remarked in chapter 1, there are a number of problems with this theory-driven picture of mindreading. An important question is how we acquire the background knowledge needed to sensitively apply our folk psychological theory in the large variety of practical contexts, without having to claim that all this knowledge is simply innate. Another

[46] See also Botterill (1996), who claims that ‘if belief-desire psychology has a central principle, it must link belief, desire and behavior. It could be formulated like this: [Action Principle] An agent will act in such a way as to satisfy, or at least to increase the likelihood of satisfaction, his/her current strongest desire in light of his/her beliefs’ (p.115).

pressing question is how we acquire the theory itself. However, there are also problems with the idea that we understand others by means of an *inferential procedure*. Wittgenstein (1953), for example, argues that ‘I know that a person who behaves in a particular way - who, for example, gets red in the face, shouts, gesticulates, speaks vehemently, and so forth - is angry precisely because I have learned the concept "anger" by reference to such behavioral criteria. There is no inference involved here. I do not reason "he behaves in this way, therefore he is angry" - rather "behaving in this way" is part of what it is to be angry and it does not occur to any sane person to question whether the individual who acts in this way is conscious or has a mental life’ (§303). Wittgenstein’s point is that our knowledge of the other mind is not primarily inferential in nature, but rather determined by public criteria that govern the application of psychological concepts. Inference seems only required under the Cartesian assumption that we have to work ‘outwards’ from the interiority of our own mind, to abstract from our own cases to the ‘internal’ world of others. This argument fits nicely with Gallagher’s (2004) observation that there is no phenomenological evidence for the claim that we use inferential principles when we are interacting with other persons. (cf. chapter 1.3)

Proponents of ST might try to avoid these problems by proposing that mindreading is *process-driven*. We are capable of accurately simulating another person as long as (i) the process driving the simulation of the other is the same as the process that drives our own system, and (ii) our initial mental states are the same as those of the other person.[47] These requirements are representative for the analogical element that is characteristic for the argument of analogy - that the other is ‘like me’ in the relevant aspects. Since the simulator and its target are probably not exactly psychologically alike, we need to feed pretend inputs into the relevant psychological mechanisms in order to come up with decent predictions and explanations. This allows us to make ‘adjustments for the relevant differences’.

The assumption of analogy allows us to understand others without theory. As Goldman (2006) puts it, ‘to read the mind of others, they need not consult a special chapter on human psychology, containing a theory about the human decision-making mechanism. Because they have one of those mechanisms themselves, they can simply run their mechanism on the pretend input appropriate to the target's initial position. When

[47] However, as Fuller (1995) points out, this also implies that simulation routines still depend on ‘a general premise stating that the model is relevantly similar to the [thing modeled]’ (p.22).

the mechanism spits out a decisional output, they can use the output to predict the target's decision. In other words, mindreaders use their own minds to 'mirror' or 'mimic' the minds of others’ (p.20). However, we noticed in the previous chapter that this makes simulation very vulnerable to a collapse into tacit theory. Mirroring processes still seem to require that ‘some elements inside the attributor causally mediate between his explicit premises and conclusions, and that the causal structure of these elements mirrors the logical structure of psychological theory’ (p.33). And this means that we cannot employ simulation routines without the help of some kind of inferential principle that enables us to reliably infer the logical conclusion from the general premise that the other is ‘like me’ and the other ‘pretend’ premises.

Another problem with the appeal to inference is that it comes with a severe *developmental constraint*. In order to infer the mental states of others, be it by means of a folk psychological theory or on the basis of an analogical premise, I already need to have some (mastery of) mental concepts. As Hutto (2004) points out, the inferential procedures employed by TT and ST make use of rather sophisticated abstract concepts such as: ‘agent’, ‘rational thinker’, ‘belief’ and ‘desire’. It remains doubtful whether, let’s say, four-year-olds, already have a handle on these concepts. One might try to sidestep this requirement by arguing that we should think of the relevant inferential processes as taking place at the *sub-personal* level, that is, in the brain. Goldman (2006), as we saw in the previous chapter, argues that mindreading is executed at the personal level by simulation, which is in turn implemented at the sub-personal level by a set of inferential principles. Simulation routines are executed by an ‘algorithm’ that is ‘tacitly represented at some level in the brain’ (p.33) The problem is, however, that such an algorithm still needs to operate upon *mental content* if we want to maintain that it functions like an *inferential* argument. And this in turn requires a sensible notion of content.

This is primarily a concern for simulation theorists who employ a *broad* notion of simulation as being essentially a mindreading process. Currie and Ravenscroft (2003) note that ‘simulation, as it is currently used, is ambiguous; it has a narrower and a broader meaning. Suppose I try to predict your behavior by imagining myself in your situation. There are three things that must go on if I am to get the answer by simulation. The first is to acquire knowledge, or at least some beliefs, about your situation. The second thing is for me to place myself, in imagination, in that situation and to see, what, in imagination, I decide. The third is to draw a conclusion from this about what you will do. Sometimes "simulation" refers to the whole three-tier process, sometimes just to the bit in the middle' (p.54). Simulation theorists who adopt a broad understanding of simulation construe it as involving an inferential procedure that follows the steps of a logical argument, and are therefore not really different from theory theorist.[48]

In this respect, it is far less demanding to articulate a narrower notion of simulation. Gordon (1995), for example, argues that simulation is not a process of *transportation* but rather one of *transformation*. This is a 'hot' methodology because it involves the exploitation of one's own motivational and emotional resources.[49] Crucially, such an imaginative transformation does not require, as he puts it, any 'inference from me to you'. In proposing his radical kind of simulation, Gordon rejects the assumption that our social encounters mainly take place against the backdrop of strong first/third-person divide. He points out that the mirror neuron processes constitutive for what Gallese calls the 'shared manifold' or 'we-space' implicitly express the *similarity* of self and other rather than their *distinctness*. They show us how the other's observed behavior and the self's matching response become intelligible *together*, that is, in the same process. When we engage in social interaction, it is not necessary for to us make any assumptions about our similarity to them, implicit or otherwise. Gordon (2005) suggests that we do not infer from the first to the third-person, but rather 'multiply the first person'.

The question is to which extent mirror neuron processes can still be interpreted as instances of simulation (cf. chapter 2.3). This is a problem for all simulation theorists insofar they hold that simulation operates 'primarily at the sub-verbal level' (Gordon 1986, p.170) and claim that, for a large part, simulation is 'non-conscious or minimally conscious'

[48] Of course, this does not mean that both positions are vulnerable to the same objections. For example, an objection against the formal validity of the argument from analogy is that it only enables me to understand myself in the situation - I don't understand the other. Wittgenstein (1953), for example, observes that 'If one has to imagine someone else's pain on the model of one's own, this is none too easy a thing to do: for I have to imagine pain which I do not feel on the model of the pain which I do feel. That is, what I have to do is not simply to make a transition in imagination from one place of pain to another. As, from pain in the hand to pain in the arm. For I am not to imagine that I feel pain in some region of his body (Which would also be possible)' (§302). Although this argument has some force against ST, it cannot be used against TT. This is because TT is committed to inference but not to analogy.

[49] See Gordon (1992), where he writes that 'In seeking an explanation of your friend's action, you were looking for features of the environment (features you believed it to possess) that were menacing, frightening, attractive, and the like. This is not a matter of looking dispassionately for features believed to produce certain characteristic actions or emotions. Rather, it is a search that essentially engages your own practical and emotional responses' (p.15).

(Goldman 2006, p.151). But those who employ a narrow notion of simulation at least have the benefit of *parsimony*, in the sense that they do not have to postulate tacit inferential principles and (non-)conceptual mental contents in order to explain our basic intersubjective engagements. Hutto (2004) argues that the tendency of simulation theorists (i.e., those who employ a broad notion of simulation) to do this in fact reveals a *theoretical bias* in their view of intersubjectivity, and he warns against the assumption that '[...] the processes involved in basic acts of recognition, even intersubjective ones, tacitly mimic those of mature reasoners who would tackle the same problem using a set of abstract concepts and general principles so as to make explicit inferences. We are systematically misled on this score because in the very act of classifying such behavior we must employ our own conceptual scheme of reference. But it is nothing more than an intellectual bias to suppose that, for example, young children or animals must be tacitly employing it' (p.557). According to Hutto, this intellectual bias is particularly hard to resist as long as it is assumed that we always start from a detached point of view in our dealings with others. The question is precisely whether such a viewpoint does justice to our everyday social engagements with others. Most of the time, we already know what to expect from others and they know what to expect from us. We do not need any mediating knowledge or inferential principles because 'much of the work of understanding one another in day-to-day interactions is not really done by us at all, explicitly or implicitly. The work is done and carried by the world, embedded in the norms and routines that structure such interactions' (McGeer 2001, p.119).

Mental concept mastery

Both the ability to introspect my own mental states and the ability to make appropriate inferences over them presuppose a certain level of mental concept mastery. To introspect a specific sensation or to infer that someone is in pain, I need to know what the mental concepts 'pain' and 'sensation' mean. But how do we acquire this knowledge? Some proponents of ST suggest that these terms get their meaning by 'inner ostension' - by being directly associated with a specific quality of internal and privately experienced mental states. This is the view of Meltzoff (2002), for example, who is often interpreted as a theory theorist, but in this respect defends a simulation approach. Meltzoff proposes that our understanding of mental states develops as follows: ‘As infants perform particular bodily acts they have certain mental experiences. Behaviors are regularly related to mental states. For example, when infants produce certain emotional expressions and bodily activities, such as smiling or struggling to obtain a toy, they also experience their own mental states. Infants register this systematic relation between their own behavior and underlying mental states’ (p.35). In a further step, infants use these ‘behavior-mental states mappings’ to make inferences about the mental states of others on the basis of an analogy.

The most fundamental flaw in proposals like these is precisely the assumption of inner ostension, i.e. that one learns from one’s own case what thinking, feeling, sensation are. Wittgenstein (1953) already showed how this leads first to solipsism, and then to nonsense. He illustrated the difficulty of inner ostension by scrutinizing the following quote from Augustine: ‘When they (my elders) named some object, and accordingly moved towards something, I saw this and I grasped that the thing was called by the sound they uttered when they meant to point it out. Their intention was shown by their bodily movements, as it were the natural language of all peoples: the expression of the face, the play of the eyes, the movement of other parts of the body, and the tone of the voice which expresses our state of mind in seeking, having, rejecting, or avoiding something. Thus, as I heard words repeatedly used in their proper places in various sentences, I gradually learnt to understand what objects they signified; and after I trained my mouth to form these signs, I used them to express my own desires’ (Confessions I 8).

The above passage indicates that Augustine assumes that language learning occurs through ostensive definition, i.e., that the meaning of a term is learned by pointing out examples. But Wittgenstein argues that this assumption is very problematic. One problem of learning by ostensive definition is that this by itself can never fix the meaning of a word. ‘No one can ostensively define a proper name, the name of a color, the name of a material, a numeral, the name of a point of the compass and so on. The definition of the number two, "That is called 'two'" - pointing to two nuts - is perfectly exact. But how can two be defined like that? The person one gives the definition to doesn't know what one wants to call "two"; he will suppose that "two" is the name given to *this* group of nuts! He *may* suppose this; but perhaps he does not. He might make the opposite mistake; when I want to assign a name to this group of nuts, he might understand it as a numeral. And he might equally well take the name of a person, of which I give an ostensive definition, as that of a color, of a race, or even of a point of the compass. That is to say: an ostensive definition can be variously interpreted in *every* case' (1953, §28).

In order to learn from ostensive definition, the learner already needs to have some grasp of what the teacher intends when pointing to something. An everyday ostensive definition is embedded in a public language, and in a social community in which that language is used. '[...] the ostensive definition explains the use - the meaning - of the word when the overall role of the word in language is clear. Thus if I know that someone means to explain a color-word to me the ostensive definition "That is called sepia" will help me to understand the word [...] One has already to know (or be able to do) something in order to be capable of asking a thing's name' (§30).

For Augustine, by contrast, language 'expresses our state of mind', and he seems to assume that language learning is essentially of matter of *understanding*. According to Wittgenstein, Augustine's account of how we learn our first language actually resembles how we learn a *second* (foreign) language (cf. §32). Learning by ostensive definition seems to imply the translation of an inner private language into an outer conventional language. The terms of this inner private language get their meaning through inner ostension, by being directly associated with a specific quality of privately experienced mental states - independently of a public language.

However, Wittgenstein gives us a powerful argument for the *impossibility* of such a private language. This is how it goes: suppose that at a certain point in time, you decide to endow the term W with meaning, solely by associating it with a certain sensation you feel at that time. At a later time, upon feeling the same sensation, you say: 'Hey, there is another W.' But how can you determine whether you have used the term correctly on this occasion? Perhaps you misremembered the fist sensation. Or perhaps you saw a close similarity where in fact there was none. In order to distinguish between the correct use of the term W and the incorrect use of W, one must have a *criterion for identification*. This entails that one must be able to follow a rule privately in isolation from others. But this is impossible, according to Wittgenstein, because *seeming* to follow a rule can never be tantamount to actually following that rule. Whatever *seems* right will *be* right, which only means that here we can't talk about right (cf. §258).

The private language argument is obviously problematic for ST insofar as the latter assumes that we first learn to identify and self-attribute mental states *in private,* and then use this as a starting point for our knowledge of the other mind. However, it is often objected that the argument draws a stronger conclusion than its premises justify. For if a public check on a correct application is all what is required for meaningfulness, then all one's understanding of 'W' need include is some connections between the occurrence of this sensation and the occurrence of other phenomena. But these other phenomena need not be publicly observable phenomena per se; they can be other mental states and still serve as checks on the correct application of 'W'. This idea is at the core of TT, according to which the meaning of folk psychological concepts such as beliefs and desire depends on their role in a larger theoretical framework. Meaning is not just given, but created as a function of prediction and explanation. And this is not necessarily a public process.

Consider Fodor's (1979) 'Language of Thought', for example. According to this MTT proposal, humans are born with a content-processing system built into their nervous system, which resembles the machine language that is hard-wired into a computer. But instead of computing binary code, this system processes mental representations by means of specific rules. These representations have contents by virtue of their ability to correspond with (things in) the world. At the same time, they are sensitive to computational processing due to their 'lingual' nature. Fodor (1979) explicitly defends the Augustinian idea that learning a first language is a translation process from an inner language to an outer language. He argues that the language of thought is what eventually enables us to learn our 'natural' language: 'You cannot learn a language whose terms express semantic properties not expressed by the terms of some language you are already able to use' (p.61). One possible objection against the language of thought is that it leads to a regress: if we cannot learn a language unless we already have one, we also need another language in order to learn the first one, and so on ad infinitum. But Fodor dismisses this objection by saying that in order to use a language, you don't need to *learn* a language – you need to *know* it. And the language of thought is known but not learned, since each of us is simply *born* with it.[50]

What about the private language argument? Fodor's language of thought is private in the sense that it is not being governed by public conventions. However, Fodor argues, this is not necessary a problem as long as it is employed in a tacit way. According to the private language argument, we are not able to follow a rule privately in isolation from others. But what if all human beings are born with the same set of folk psychological rules?

[50] Notice that such an appeal to innateness is typical for the rationalist approach to the problem of knowledge (cf. chapter 1.2).

What if we are natural born rule-followers? Again, the appeal to tacit theory is tempting. However, we already saw in chapter 1 that the assumption of a tacit set of (innately acquired) theoretical principles still does not solve the problem of how we are able to sensitively apply our mindreading skills in a large range of practical contexts. It seems that we need more than simple belief-desire syllogisms in order to select the specific contents of the mental states over which our folk psychological theory quantifies in particular situations. How do we acquire the background knowledge needed to pull this off? Fodor has only one answer to this question: innateness. But if everything is already in place before we acquire our 'natural' linguistic skills, another question pops up. How can we make sense of the mental content that is needed to fuel our tacit theory of mind? It is notoriously difficult to spell out what is precisely meant by a tacit belief or desire. This is precisely why some have suggested that, when it comes to specifying the content of the belief and desires of nonverbals, folk psychology increasingly comes 'under stress' (cf. Godfrey-Smith 2003). The root problem, according to Hutto (2007a), is that the very idea of content as something 'given' in perceptual encounters, 'acquired' by mental states, and 'manipulated' in sub-personal cognitive processing is deeply problematic if not incoherent. Remark that the notion of mental content is not only problematic for modular TT (MTT), but also for scientific TT (STT). Although STT rejects the claim that mental content is innate, it still needs to explain how our folk psychological principles facilitate the acquisition of non-conceptual mental content during development.[51]

A question of analogy

So far I have not dealt with something that is crucial to the argument of analogy. This is the idea of analogy *itself*. My evaluation of this requirement entirely depends on how it is

[51] Many proponents of TT think that the basic perceptual acts of nonverbal creatures are content-involving. Evans (1982) gives us an adequate description of what this means: 'In general, we may regard a perceptual experience as an informational state of the subject: it has a certain *content* -- the world is represented a certain way - and hence it permits of a non-derivative classification as *true* or *false.* For an internal state to be so regarded, it must have appropriate connections with behavior - it must have a certain motive force upon the actions of the subject [...] The informational states which a subject acquires through perception are *non-conceptual*, or *nonconceptualised.* Judgements *based upon* such states necessarily involve conceptualisation' (pp.226-7). The idea is that perception involves a translation (or conceptualization) of the contents of one language into another.

interpreted. As long as we conceive of analogy in terms of a (tacit) premise, which states that the interpreter is similar to the agent under consideration and serves as a starting point for a (conscious) inferential procedure, I think the requirement of analogy is very problematic. Such an interpretation is usually endorsed by simulation theorists that employ a broad notion of simulation. It not only introduces a number of developmental constraints (most importantly, that of mental concept mastery), but is also vulnerable to a number of standard objections against the *argument* from analogy. (cf. chapter 2.2) But what if it is possible to have the argument from analogy without the actual argument? What if there is an analogy between ourselves and others that is non-conceptual and non-inferential in nature?

Gallagher (2003a) points out that such an analogy, a kind of 'common code', may be found at the level of sensory-motor mechanisms. He claims that developmental studies suggest that this common code is already operative from the very beginning of life: 'What I see is automatically registered in a code that is common to other sense modalities, including proprioception; and in the case of seeing biological movement, perception includes motoric, kinaesthetic activation. So when I see the other's body moving in a certain way, I have a kinesthetic-proprioceptive sense of what that is like.' Analogy, thus understood, seems to be far less demanding. And it can even be used against those who are critical of the argument from analogy. For example, Zahavi (2001) argues that for the argument from analogy to work, there has to be a similarity between the way in which my own body is given to me, and the way in which the body of the other is given to me. But Zahavi points out that my own body, as it is *felt* proprioceptively for me, does not at all resemble the other's body as it is *perceived* visually by me. However, if we can find a non-conceptual and non-inferential analogy between ourselves and others at the sub-personal level, this argument appears to be off-base in an important way. It is tempting to argue that such an analogy would in fact prove that we are born with some kind of inner language of thought. This would be a mistake, however, since it would be a very strange language – a language of which we are not conscious, and which does not involve inference, concepts or content. In other words, this sub-personal language would lack all the important properties we normally attribute to language.

3.4 Beyond the problem of the other mind

Given that our introspective and inferential abilities presuppose a rather sophisticated knowledge of mental concepts, the bottom line question is how we acquire these concepts and come to learn what they mean. Many proponents of TT and ST think that this is the achievement of the *individual agent.* Meaning is primarily a private affair – it is 'given' through introspection (Goldman's ST), explained in terms of innateness (modular TT) or picked up from the environment through perception (scientific TT).[52] With such a narrow Cartesian focus on *subjectivity*, it seems almost natural to assume that intersubjectivity is *derivative* – a matching process between individualized mental states that share the same meaning. It also seems inevitable that, on such a view, intersubjectivity turns out to be very problematic. As long as we are inspired by a Cartesian phenomenology of uncertainty, the pressing issue remains how we can access other minds, i.e. what sort of intervention process (inference, introspection, or analogy) is needed between our initial perception of others and our active response towards them.

TT and ST assume that in order to solve this problem, we have to start with the primacy of a theoretical, third-person stance towards others. But a much more basic question is whether the Cartesian context is the primary context in which intersubjectivity takes place. Gallagher (2001) is right to stress that it is questionable whether our ordinary attempts to understand other people are best characterized as explanations and predictions. Most of our intersubjective encounters are firmly rooted in *second-person interactions*, in which we directly engage with others and already know to some extent what we can expect from them. Of course, there are situations in which we can be perplexed by their actions, and try to predict their next move or explain what exactly motivated them to behave in a certain way. As Hutto (2007a) points out, 'driven by suspicion we may be left with nothing but speculation and supposition about their motives. That is, we may be forced to make third-party predictions and explanations of actions precisely in the sorts of cases in which we do not know what to expect from others or when we cannot engage with them directly. But, for this very reason, these sorts of approaches

[52] Externalist TT (cf. chapter 1.2) is clearly the exception here, and I fully agree with their objection to internalist versions of TT that folk psychological principles 'ain't in the head'. At the same time, however, externalist TT still takes the prediction/explanation of behavior by means of a folk psychological theory to be central to intersubjectivity, and in this respect remains firmly rooted in the Cartesian tradition discussed in the previous sections.

are bound to be, on the whole, much less reliable than our second-person modes of interaction' (p.13).

The moral is that, in practice, there is no *general* problem of the other mind. Why should we assume that intersubjectivity is intrinsically problematic and best characterized in terms of a phenomenology of uncertainty? If we pay attention to practice, we find that most of the time we already have some basic understanding of what to expect from others, and we also know what they expect from us. We do not need to engage in inferential or introspective procedures to make sense of what they mean or what they are doing. Gallagher (2001) claims that 'before we are in a position to theorize, simulate, explain or predict mental states in others, we are already in a position to interact with and to understand others in terms of their gestures, intentions and emotions, and in terms of what they see, what they do or pretend' (p.91).

Importantly, these interactions provide us with a basic understanding of other minds that is not subject to *reasonable* doubt. This is well expressed by Thomas Reid (1983), who was much more practically minded than John Stuart Mill in this respect. Reid dismissed the problem of the other mind by arguing that 'No Man thinks of asking himself what reason he has to believe that his neighbor is a living creature. He would be not a little surprised if another person should ask him so absurd a question: and perhaps could not give any reason which would not equally prove a watch or a puppet to be a living creature. But, though you should satisfy him of the weakness of the reasons he gives for his belief, you cannot make him in the least doubtful. This belief stands upon another foundation than that of reasoning and therefore, whether a man can give good reasons for it or not, it is not in his power to shake it off' (pp.278-9). Reid is right that the problem of the other mind does not show up in our common-sense encounters with others. But of course, we might wonder what to make of this foundation that secures our understanding of the other mind. How do we interact with others 'in terms of their gestures, intentions and emotions'?

In the next chapter, I attempt to answer the above questions by making a case for the importance of *second-person practices*. These intersubjective engagements embody our baseline understanding of others, and enable us to relate to them in a direct way - without mindreading or other cognitive/conceptual interventions. Such a pragmatic approach is able to avoid the severe developmental constraints that need to be met by TT and (most) ST accounts, and gives us more insight in the context-sensitivity of our intersubjective capacities. Moreover, it also does more justice to the empirical evidence on their *actual* *development*. From a pragmatic point of view, this is one of the first requirements that a plausible account of intersubjectivity has to satisfy. And last but not least, my pragmatic approach allows for a richer *phenomenology* of intersubjectivity - one that does not need to be characterized solely by means of prediction and/or explanation. This is because many of the anticipatory and predictive processes that enable our meetings with other minds take place at the neurobiological level, and can be described in *sub-personal* terms.

4.

Mind Shaping in Early Ontogeny

That many operations of the mind have their natural signs in the countenance, voice and gesture, I suppose every man will admit. The only question is, whether we understand the significations of those signs, by the constitutions of our nature, by a kind of natural perception similar to the perceptions of sense; or whether we gradually learn the signification of such signs from experience, as we learn that smoke is a sign of fire or that freezing is a sign of cold [...] It seems to me incredible, that the notions men have of the expressions of features, voice, and gesture, are entirely the fruit of experience.

- Reid 1983

The mind in action

The previous chapters mainly dealt with intersubjectivity through the theory-colored spectacles of TT and ST. Consequently, we have primarily focused on social encounters in which agents were portrayed as bystanders, merely observing others without actively interacting with them. In such a context, intersubjectivity is primarily about mental state management. The mind is presented as an autonomous spectator, and knowledge of the other mind is considered to be one of its cognitive and conceptual achievements. The body is supposed to facilitate this process, but it is not supposed to play a *constitutive* role.

My own approach, by contrast, is firmly rooted in the pragmatist assumption that the mind is fundamentally shaped by its bodily existence (embodiment) and cannot be understood in isolation from its environment (embedment). It borrows from *enactivism* insofar it subscribes to a conception of the mind as emerging from the intricate web of interactive processes that is characteristic for a *complex system*. Complex systems are self-generating and self-maintaining wholes, which define their boundaries through their interaction with the surrounding world (cf. Varela 1979, Thompson 2007). A system is complex in virtue of the dynamic processes that hold between its sub-systems, and this is why its (emergent) properties cannot be fully explained in terms of these sub-systems alone (cf. Cilliers 2005). In order to understand a complex system, it is necessary to take into account the various interactive processes that describe its organization and define it as a system. In order to understand the complex system that is mind, we must pay attention to the dynamic processes between brain, body and environment that give rise to it. At the same time, however, the mind is more than a coupled system of brain, body and environment in isolation. The mind is stimulated, constrained and co-constituted by *other* coupled systems, and emerges as the result of continuing interactions with *other minds*.

This chapter shows how, at a very basic level and without cognitive and/or conceptual requirements, such interactions can be explained in terms of *second-person practices* (see fig. 4.1).[53]

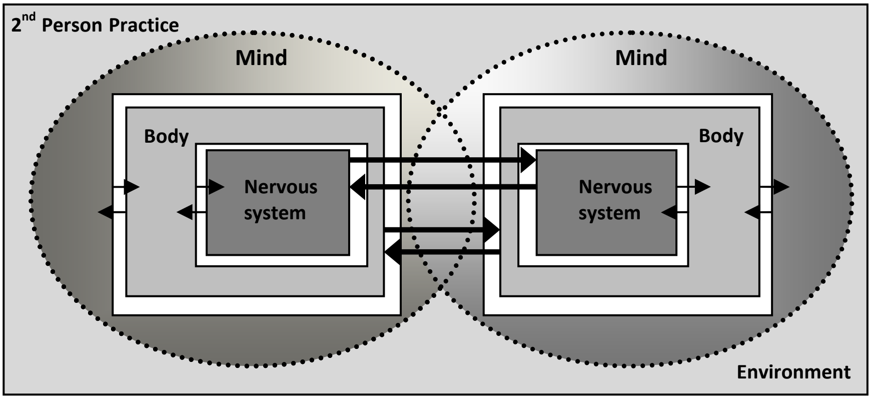


Fig. 4.1 Interacting minds in a second-person practice. Minds dynamically 'co-emerge' as the result of a constant interaction between nervous system, body and environment

[53] I share this starting point with many other enactive approaches to intersubjectivity (e.g., Fuchs and De Jaegher 2009, Gallagher and Zahavi 2008, Hutto 2007, Iacoboni 2003, Ratcliffe 2007, Thompson 2007).

These embodied and embedded ways of dealing with others constitute the base-line for social understanding, and they provide the background knowledge required for our more sophisticated modes of intersubjectivity. There are two ways in which these practices are primary to more advanced forms of social understanding. In the first place, they involve social abilities that come *earlier* in development and may even be partially *innate*. Secondly, they are also primary in the sense that they *continue* to characterize most of our social interactions throughout ontogeny, and remain the *default* mode of how we understand others.

The first part of this chapter shows that many embodied practices are already up and running from the moment we are born. I start by discussing a broad range of empirical findings demonstrating that very young infants are already able to interact with others in a rather sophisticated way.[54] Empirical research on early imitation reveals that neonates manifest a very primitive form of *co-consciousness*, in the sense that they have a proprioceptive awareness of both self and other. During the first year, various embodied practices trigger the infant to develop this awareness into a more advanced action-based understanding of intentional and emotional behavior (section 1). These practices are not self-sufficient. They depend on and are shaped by our bodily existence and various (partly) innate sensory-motor capacities (section 2). At around one year, infants acquire abilities that allow for a more advanced understanding of others in terms of their involvement in pragmatic contexts (section 3). The defining feature of these embedded practices is, as Hobson (2002) puts it, that 'an object or event can become a focus between people. Objects and events can be communicated about […] the infant's interactions with another person begin to have reference to the things that surround them' (p.62).[55] Altogether, these practices provide infants by the end of the second year with a large body of pre-theoretical knowledge - the 'know how' required for the more advanced (narrative) modes of intersubjectivity that will be discussed in chapter 5.

[54] Some of the empirical evidence that is reviewed in this chapter is also put forward to support TT and/or ST approaches to intersubjectivity. However, I aim to show that it fits more comfortably with a pragmatic story about intersubjectivity, since such a story takes their functioning at face value and looks at what infants are *actually doing in practice*, as supposed to hypothesizing what should be going on in theoretical or simulation terms.

[55] I call these practices 'embedded practices' because they allow for a more advanced, 'situated' form of social understanding.

4.1 Embodied practices [#]

Early sympathizers

By the time we are born our capacities for intersubjectivity are already shaped by our body and its movement. Bodily movement, as Gallagher (2005) aptly puts it, has already been organized in proprioceptive and cross-modal registrations in order to provide the capacity for differentiation between self and non-self. 'Movement and the registration of that movement in a developing proprioceptive system contributes to the self-organizing development of neuronal structures responsible not only for motor action, but for the way we come to be conscious of ourselves, to communicate with others, and to live in the surrounding world' (p.1).

Developmental studies point out that neonates indeed manifest a clear sense of self as a differentiated and situated entity in the world. Rochat and Hespos (1997), for example, have shown that they are already capable of discriminating between external and self-stimulation. In the external stimulation condition of their study, the index finger of the experimenter touched one of the infant's cheeks. In the self-stimulation condition, the infants spontaneously brought one hand to their face, touching one of their cheeks. The study revealed that neonates displayed significantly more rooting responses (i.e., head turn towards the stimulation with mouth open and tonguing) following external stimulation compared to self-stimulation. Neonates are not only able to discriminate between themselves and their environment, but they also respond selectively to other human agents. Despite not yet having acquired the appropriate concept of 'agent' or 'face', they differentiate effectively between agents and non-agents, and faces and non-faces.

It has been shown that very young infants are particularly sensitive to the *emotions* of other people, expressing what Trevarthen (1979) called 'intersubjective sympathy'. For example, Field et al. (1982) have shown that, as soon as 36 hours after their birth, neonates are already capable of discriminating the facial expressions happy, sad, and surprised. They also produce much more reactive crying when they hear the sound of another neonate crying instead of white noise or a synthetic cry (cf. Sagi and Hoffman 1976, Martin and Clark 1987).[56]

[#] Section 4.1 has been written in collaboration with Sanneke de Haan, and I want to acknowledge her for several insights presented here.

A good illustration of the infants' responsiveness to the emotions of others is *affective synchrony*, which begins to occur in mother-infant interactions when infants are around 2-3 months of age (Stern 1985, Trevarthen 1979). Both mother and infant contribute to these affect-sharing episodes, using an increasing repertoire of interactive behaviors. A closer look at these specific social interactions (so-called 'microanalyses') reveals that mothers are highly likely to imitate infant expressions of enjoyment and interest, as well as expressions of surprise, sadness, and anger (Malatesta and Haviland 1982). However, they rarely display negative emotions to their infants. Infant-mother interactions exhibit considerable positive synchrony, partly as a consequence of the mother's contingent matching of positive infant emotional expressions.[57]

Stern (1985) claims that the early interactions between infants and their caregivers are first and foremost directed at the *attunement* of affect. He coins the term 'vitality affect' to clarify how different modalities can have the same 'kinematics' and thus express the same affect. For example, a mother can sooth her baby by saying 'there, there' in a comforting tone of voice, or by re-assuringly stroking the baby's back. The rhythm of speaking and the rhythm of stroking are the same, and in both allow the mother to express the vitality affect of soothing.

Stern emphasizes that we need more than imitation alone to explain what happens in such interactive exchanges.[58] He also notes that the first interactions between infants and caregivers typically entail matching the same vitality affect in the same modality, whereas from roughly 9 months on, caregivers are more inclined to react with the same vitality affect in a *different* modality. However, there is evidence that 5-month-old infants are

[56] What is interesting about this example is that neonates do not seem to respond to the sound of their *own* cries (on audiotapes). This supports the claim that there already is some kind of self-other distinction functioning right from birth.

[57] But this also works in the opposite direction. For example, Field et al. (1985) documented how depressed mothers influence their infants through these interactions in a negative way.

[58] Stern (1985) writes: 'For there to be an intersubjective exchange about affect, then, strict imitation alone won't do. In fact, several processes must take place. First, the parent must be able to read the infant's feeling state from the infant's overt behavior. Second, the parent must perform some behavior that is not a strict imitation but nonetheless corresponds in some way to the infant's overt behavior. Third, the infant must be able to read this corresponding parental response as having to do with the infant's own original feeling experience and not just imitating the infant's behavior' (p.139). The mere reproduction of the other's over behavior does not yet give us a clue that the other person really has a similar experience. It is exactly the slight modulation, for instance a change in the modality of expression that reveals the idiosyncrasy of the other and the individuality of their expression.

already able to detect a correspondence between different modalities that specify the expression of an emotion, such as visual and auditory information (Walker 1982; Hobson 1993, 2002). In any case, what is important here is that there appears to be a growing differentiation and complexity in the affect attunement of young infants. As Gopnik and Meltzoff (1997) put it, they increasingly interact with others in ‘a way that seems “tuned” to the vocalizations and gestures of the other person’ (p.131).

Early responders

From very early on children already show responsiveness to goal-directed or intentional behavior.[59] A series of experiments by Leslie (1982, 1988), for example, indicates that by 5 months, infants perceive intentionality and have different expectations about the effects on another object of the actions of a human hand versus an inanimate object. Woodward (1998) agrees. By habituating 5-month-old infants to a hand reaching for one of two objects, she found that they looked longer when the hand reached for the object not previously obtained, regardless of its position. She concluded that the infants were not ‘encoding’ the structural elements of the display (e.g., movement to the left or to the right), but the *goal* of the actor’s reach. This was further supported by a condition where the infants did not look longer when the hand was replaced by a metal rod (which helped to rule out an explanation in terms of a conditioned response, or at least one formed during the habituation phase). By 9 months, infants are able to follow the other person’s eyes and start to perceive various movements of the head, the mouth, the hands, and more general body movements as meaningful, intentional movements (Senju et al. 2006). And at around 10 months, infants have learned to parse specific kinds of continuous action according to intentional boundaries (Baird and Baldwin 2001, Baldwin et al. 2001).

Baron-Cohen (1995) has proposed to explain this early responsiveness to intentional action in terms of what he calls an ‘intentionality detector’ (ID): a perceptual device that

[59] My use of the term ‘intentional’ here is in line with Hutto’s (2007) description of ‘intentional attitudes’. According to Hutto, preverbal infants display intentional attitudes insofar as they selectively respond to certain aspects of their environment. However, intentional attitudes should not be confused with *propositional* attitudes. The latter are exclusively employed by those beings that have mastered certain linguistic constructions and practices, including the ability to represent and reason about complex states of affairs in truth-evaluable ways.

enables neonates to distinguish animate from inanimate objects. He argues that the ID is activated 'whenever there is any perceptual input that might identify something as an agent [...] This could be anything with self-propelled motion. Thus, a person, a butterfly, a billiard ball, a cat a cloud, a hand, or a unicorn would do' (p.33). The ID is supposed to be a kind of device that allows the infant to read 'mental states in behavior' by interpreting 'motion stimuli in terms of the primitive volitional mental states of goal and desire' (p.32). Baron-Cohen thinks that goals and desires are *primitive* mental states because they are minimally required to make sense of the universal movement of all animals: approach and avoidance. This is how he puts it: 'If you see an animal moving, be it an amoeba, a mouse, or a British prime minister, all you need to refer to in order to begin to interpret its movement are these two basic mental states' (ibid.).

However, as I already pointed out in previous chapters (cf. chapter 1.3 and 2.1), there are serious problems with the idea of locating mental states at the sub-personal level. Moreover, the question is whether it is *necessary* to do so. Do we really need to postulate primitive mental states such as desires and goals in order to make sense of the infants' responsiveness to intentional action? Gallagher (2001) thinks not. He suggests that the ID allows the infant to perceive intentional movement in a non-mentalistic way, and approvingly cites Scholl and Tremoulet (2000), who claim that the ID is 'fast, automatic, irresistible and highly stimulus-driven' (p.299).

A similar, but somewhat more advanced version of the ID is what Baron-Cohen (1995) calls the 'eye-direction detector' (EDD). The EDD is more specific than the ID since it is linked directly to the perception of faces, in particular the eyes. According to Baron-Cohen, the first function of the EDD consists of the detection of eye-like stimuli. Whenever the EDD detects eye-like stimuli, it 'fixates on these for relatively long bursts and starts to monitor what the eyes do' (p.39). The EDD builds on the idea that young infants already have a natural preference for looking at the eyes of other persons over looking at other parts of their face. For example, it has been shown that, at the age of 2 months, infants look almost as long at the eyes as at the whole face, but significantly less at other parts of the face (cf. Hainline 1978; Maurer and Barrera 1981, 1985).

Baron-Cohen suggests that the EDD has a second function as well: it enables the infant to determine whether the eyes it is looking at are directed at itself or at something else. There is some evidence that infants are already able to do this at a very young age. For example, it has been shown that 6-month-old children look approximately two and a half minutes longer at a face looking at them than at a face looking away (Butterworth 1991, Vicera and Johnson 1995).

The third function of the EDD, according to Baron-Cohen, is to ‘infer from its own case that if another organism’s eyes are directed at something, then that organism sees that thing’ (1995, p.39). Such an inference is necessary in order to understand that the other person actually sees what he or she is looking at. However, Gallagher (2001) has argued that this assumption is mistaken, because it is only by virtue of e*xperience* that the infant comes to discover that someone could be looking in a certain direction without actually seeing something. This is something we learn rather than a default mode of the EDD: ‘*on the face of it*, that is, at a primary (default) level of experience, there does not seem to be an extra step between looking at something and seeing it’ (p.89, italics in original).

In a certain sense, however, this seems to be precisely what Baron-Cohen is proposing. He suggests that 'from very early on, infants presumably distinguish seeing from not-seeing [...] Although this knowledge is initially based on the infant's *own experience*, it could be generalized to an Agent by analogy with the Self' (p.43, italics added). What is problematic here is precisely the assumption that the infant comes to distinguish between seeing and not-seeing on the basis of its *own* experience, and consequently has to generalize this on the basis of an analogy. This shows that Baron-Cohen not only assumes that young infants already possess mental concepts, but also that they are able to make inferences over them on the basis of an analogy. However, as Hutto (2007a) points out, basic one-to-one interactions such as the above are not rightly characterized as involving an analogical comparison with others, or the neutral observation of outward behavior followed by cold inferences that the other is in such and such mental state. This is not only because these abilities come with severe developmental constraints, but also because there is a much more pragmatic explanation available, as we will see in a few sections.

There is also a terminological problem with Baron-Cohen’s approach. An important drawback of notions such as ‘detector’, ‘device’ and ‘mechanism’ is that they invite a *mechanical* description of what goes on during these interactions. The notion of responsiveness is much more appropriate because it emphasizes the *interactive* nature of our involvements with others. It is often taken for granted that children need to posses certain individual abilities *before* they are able to participate in embodied practices. But this assumption is problematic insofar it obscures the fact that these abilities often develop in and through the kind of interactions they are supposed to precede and explain. Therefore, the quest for the 'underlying mechanisms of change' (Striano and Reid 2006) that motivates much infant-research seems to be misguided to the extent that it is aimed at pin-pointing the individual 'pre-cursors' of our 'full-fledged' interactive abilities. Such a linear and individually centered account of the acquisition of our social know-how does no justice to the intersubjective dynamics of development, in which the mechanisms themselves are subject to dramatic change as well.

Early imitators

So far I have not paid attention to *imitation* - an ability that is crucial to infants' development, since it provides them with numerous new opportunities to explore the field of intersubjectivity. The body of research on imitation is impressive. Meltzoff and Moore (1983), for example, have shown that one hour after they are born, neonates already imitate a variety of facial gestures such as mouth-opening and tongue-protrusion. Slightly older infants, with greater neuromuscular control, can imitate more specific behaviors such as tongue protrusion to one side (Meltzoff and Moore 1995). Although their first imitative attempts lack a high degree of accuracy, infants learn to correct and improve their gestural performance over time. This allows them to increasingly fine-tune and sophisticate their interactions with others.

I should point out that the second-person interactions in which imitative behavior is embedded are better characterized in terms of embodied resonance than in terms of pure mirroring – again because of the mechanical and reflex-like connotation of these latter terms. Tomasello (1999), for instance, has suggested that young children are 'imitation machines' (p.195). However, such a mechanical view cannot explain why infants are more likely to imitate after they have been attended to by the experimenter, as Csibra and Gergely (2009) have shown in recent experiments. The notion of embodied resonance, by contrast, allows us to account for the individual modulations infants bring to bear in their interactions. They do not completely *merge* into each other, but instead mutually *tune in* to each other. Their individual modulations attest to their autonomy: for perfect contingency you only need a mirror, but for genuine social interaction you need another person.[60]

Research shows that infants from 3 months on prefer these slight modulations (e.g., time-delay) in their embodied responses, except for autistic children who continue to prefer perfect contingency (Gergely 2001). Whereas *perfect* contingency only reflects one's own agency, *imperfect* contingency suggests the influence of another person and thus interpersonal contact. Given that normal infants are still exploring their sense of agency during this period, it seems natural to assume that they are mainly interested in finding out what they *themselves* effectuate. However, as soon as their sense of agency has reached a certain level of sophistication, a pure reflection on their own deeds probably becomes a bit boring - especially compared to the novelty that is introduced by interactions with other persons. Autistic children, however, continue to prefer perfectly contingent feedback to modulated feedback. Gergely (2001, p.418) explains this in terms of the 'faulty switch' of a postulated 'contingency detection module', which leads to symptomatic difficulties in social interactions. Although there is still an ongoing debate on the underlying mechanism(s) of autism, I am skeptical whether this talk about modules will bring us any further. But given their difficulties in social interaction and problems in dealing with novelty, it is not surprising that both the suggestion of another person and the possibility of interpersonal contact are less attractive to autistic children.

Meltzoff and Moore (1994) have investigated nine characteristics of early imitation in infants under 2 months:

1. Infants imitate a range of acts
2. Imitation is specific (tongue protrusion leads to tongue not lip protrusion)
3. Literal newborns imitate
4. Infants quickly activate the appropriate body part
5. Infants correct their imitative efforts
6. Novel acts can be imitated
7. Absent targets can be imitated

[60] As De Jaegher and Di Paolo (2007) remark, participatory sense-making is only participatory as long as the participants remain autonomous. Otherwise it would be merely one person forcing a sense upon another, a one-way interaction (see also Fuchs and De Jaegher 2009).

8. Static gestures can be imitated
9. Infants recognize being imitated [61]

They point out that there is an interesting developmental change in the infants' expression of imitative behavior. Although their abilities to imitate are in place right from the off, infants still need a lot of practice to pull of the more advanced modes of imitation that come later in development. For example, neonates imitate novel acts, but research on older infants reveals a generative imitation of novelty that is beyond the scope of younger infants (Bauer and Mandler 1992, Barr et al. 1996). More in general, there seems to be a progression in imitation from pure body actions, to actions on objects, to using one object as a tool for manipulating other objects. The question is: how can we explain this progression in imitative skills?

This is where Meltzoff and Moore (1994) offer us the `active intermodal mapping' (or AIM) hypothesis (see fig. 4.2). The basic idea behind the AIM hypothesis is that imitation is essentially a 'matching-to-target' process. The active nature of this matching process is captured by a 'proprioceptive feedback loop'. The loop allows the infant's motor performance to be evaluated against the perceived target and serves as a basis for correction. This process is facilitated by a 'supramodal perceptual system' that translates visual input into motor output, and lets perception and action communicate with each other within the same 'language'. It enables the infant to recognize a structural equivalence between its own acts and the ones it sees. A successful matching between perception and action is what grounds its apprehension that the other is, in some primitive sense, 'like me'. Gopnik and Meltzoff (1997) propose to explain this intermodal and intersubjective mapping as a primitive form of *theorizing*.

[61] Notice that the imitation described in these experiments cannot be a matter of *reflex behavior* or *release mechanisms*. Reflex and release mechanisms are highly specific, and no such mechanism could exist for imitation in general. Yet the range of behaviors displayed by the infants in these studies would require the unlikely assumption of distinct release mechanisms for each kind of behavior: tongue protrusion, mouth openings, lip protrusion, head movement, finger movement, as well as smile, frown, and so forth. Importantly, the studies that show imitative behavior after a delay clearly indicate the involvement of memory. It should also be remarked that the infants improve or correct their imitative response over time. They get better at the gesture after a few practices. Neither delayed reaction nor improved performance is compatible with a simple reflex or release mechanism.

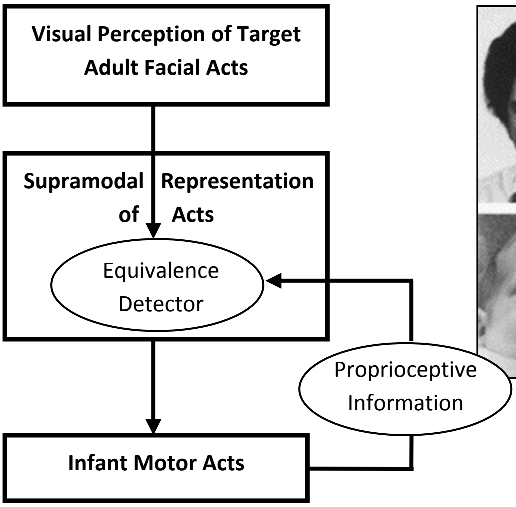

Visual Perception of Target Adult Facial Acts
Supramodal Representation of Acts
Equivalence Detector
Proprioceptive Information
Infant Motor Acts

Fig. 4.2 The AIM hypothesis

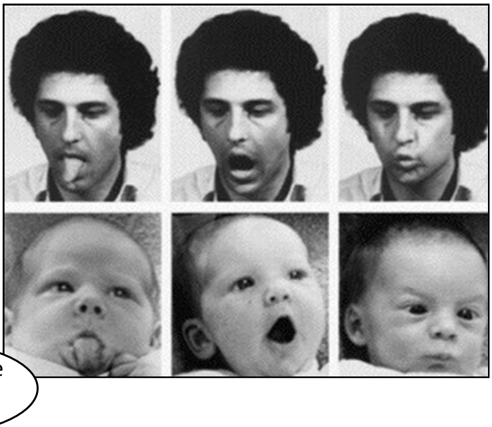


Fig. 4.3 Neonate Imitation (Meltzoff and Moore 1977)

This lies at the beginning of an inference-like operation that is eventually promoted into a theoretical attitude. Meltzoff (2002) gives us a more comprehensive description of what this implies in terms of development:

(i) Innate equivalence between self and other. Infants can imitate and recognize equivalences between observed and executed acts. This is the 'starting state', as documented by motor imitations in newborns (fig. 4.3).
(ii) Self learning. As infants perform particular actions they have certain mental experiences. Behaviors are regularly related to mental states. For example, when infants produce certain emotional expressions and bodily activities, such as smiling and struggling to obtain a toy, they also experience their own mental states. Infants register this systematic relation between their own behaviors and underlying mental states.
(iii) Others in analogy to the self. When infants see others acting similarly to them, they project that people are having the same mental experience as they themselves when performing those acts. They use the behavior-mental states mappings registered through

their own experience to make inferences about the internal states of others.[62]

Meltzoff (2002) proposes that infants gradually learn to understand others by using knowledge of how they feel when they produce an expression to infer how another feels. He argues that infants 'imbue' the acts of others with 'felt meaning', because they are able to recognize the similarities between their own acts and those of others. 'Their experience of what it feels like to perform acts provides a privileged access to people not afforded by things. It prompts infants to make special attributions to people not made to inanimate things that do not look or act like them' (p.35).

The problem is that Meltzoff's account (just like that of Baron-Cohen) presupposes all the traditional ingredients of a *mindreading* account of intersubjectivity: mental concept mastery, inferential abilities, and the analogical argument. It is highly improbable, however, that these requirements are already within the reach of young infants (cf. Bermudez 1998, Gordon 2004). Moreover, it is not clear *why* we need them to explain the basic form of social understanding that these children are capable of. As we will see in the next sections, it is very well possible to give an explanation of the matching-to-target process that underlies imitation in *sub-personal terms*, without having to refer to mindreading or mental state management.

Body image and body schema

So far I have discussed a number of embodied practices that provide young infants with a basic but effective social understanding of others. I have emphasized that these interactions should not be interpreted in terms of *mindreading*. Rather, as Hutto (2007a) claims, 'we react directly to the attitudes of others as expressed bodily and we do so because of our natural predisposition, some of which gets reformed by experience and enculturation. It cannot be stressed enough that on this model the intervening cognition that makes this possible is not fueled by representations of the behavior or mental states of others' (p.115). But of course, the important question then becomes how we can further articulate such a 'direct reaction' to the attitudes of others without appealing to

[62] See Tomasello (1999) for a similar view. Tomasello claims that 'children make the categorical judgment that others are 'like me' and so they should work like me as well' (pp.75-6).

mindreading procedures, mental concept mastery, or analogical inferences.

With respect to early imitation, the question is how to explain the fact that children are able to successfully match their *perception* of the other person with their own imitative *action*. This is even more puzzling in cases of facial imitation in which infants are not able to perceive their *own* action. Bermudez (1998) formulates the problem as follows: ‘Facial imitation involves matching a seen gesture with an unseen gesture, since in normal circumstances one is aware of one’s own face only haptically and proprioceptively. If successful facial imitation is to take place, a visual awareness of someone else’s face must be apprehended so it can be reproduced on one’s own face’ (p.125).

What is needed here is something that allows for a dynamic *co-constitution* of perception and action and explains their common coding, without requiring some kind of inferential/conceptual process to mediate between them. The problem with the proposals discussed above is the appeal to ‘internal representations’ or ‘behavior-mental states mappings’ in their explanation of such an action-perception loop. Gallagher (2005) argues that a supramodel system that integrates action and perception should not be explained in terms of 'abstract representations', but rather as a set of pragmatic (action-oriented) capabilities embodied in the developing nervous system. These capabilities constitute what he calls the *body schema*: a ‘system of sensory-motor capacities’ that functions without reflective awareness or the perceptual monitoring in an immediate and close to automatic fashion. This body schema makes it possible for children to develop a *body image*. A fully developed body image consists of a set of intentional states and dispositions such as perceptions, attitudes, and beliefs about one’s own body.[63] It involves a form of reflexive and self-referential intentionality that allows me to experience my body as ‘mine’. In case of neonate facial imitation, however, the infant does not yet possess a body of beliefs, attitudes or conceptions about its body, nor a visual perception of its own face. The only aspect of the body image available to the infant at this stage in development,

[63] Studies involving the notion of body image frequently distinguish three elements: (a) the subject's perceptual experience of his/her own body; (b) the subject's conceptual understanding of the body in general; and (c) the subject's emotional attitude toward his/her own body (cf. Cash and Brown 1987, Gardner and Moncrieff 1988, Powers et al. 1987). Although body schema and body image usually function synchronously, a few cases have been described in which one of them is dysfunctional. For instance, patients suffering from deafferentation have no proprioception from the neck down and can be said to have a defective body schema. In order to be able to move, they depend on their body image, and simple actions such as walking and holding a cup therefore require a great amount of concentration. Cases of hemi-neglect, in which patients consistently ignore one side of their body, can be interpreted as a sign of a defective body image.

according to Gallagher, is the *proprioceptive awareness* (PA) of its own body. PA is a primitive form of consciousness or pre-reflective awareness that informs the infant about the location of its limbs and its overall posture (without the aid of visual perception). Gallagher argues that this PA enables the newborn to 'know' that its own face is in some way equivalent to the visually presented face it is imitating.

More is needed to explain how PA is related to visual perception and the body schema, however. Therefore, Gallagher also puts forward the notion of *proprioceptive information* (PI), which consists of non-conscious and sub-personal, physiological information that updates the motor system about the position of body parts and movement of the body in general. Importantly, he argues that PA and PI are two sides of the *same* coin that is proprioception.[64] With this 'dual nature' of proprioception on the table, Gallagher is now able to explain how cross-modal communication between vision and proprioception is at the same time a communication between sensory and motor aspects of behavior. Since PI and PA depend on the same physiological mechanisms (the body schema), there is 'an immediate connection, a close interactive coordination, between proprioceptive information, which updates motor action at the level of the body schema, and proprioceptive awareness, as a pre-reflective, performative accompaniment to that action' (2005, p.76). And because PA and vision are intermodally linked, there is also a link between vision and PI, or more generally between sensory/perceptual and motor activities.

Early facial imitation, according to Gallagher, depends on both PA and PI. What the infant sees 'gets translated into a proprioceptive awareness of her own relevant body parts; and PI allows her to move those parts so that her proprioceptive awareness matches up to what she sees' (ibid.). But this translation is not really a translation or a transfer, because it is 'already accomplished' and 'already intersubjective'.

One of the drawbacks of Gallagher's proposal is that it promotes embodiment, but at the same time lacks in neurophysiological detail. As Edelman (1992) already made clear, 'it is not enough to say that the mind is embodied; one must say how' (p.15). It must be admitted, however, that Gallagher himself acknowledges this. He points out that 'recent studies in neuroscience suggest that there are specific neurophysiological mechanisms that can account for the intermodal connections between visual perception and motor behavior. These are mechanisms that operate prenoetically, as general conditions of

[64] I actually prefer the term *kinaesthesia* over *proprioception*, since this place a greater emphasis on *motion* instead of perception. However, to avoid confusion I will follow Gallagher in his use of the term proprioception.

possibility for motor stability and control, but are also directed related to the possibility of imitation' (2005, p.77). I will take a look at these findings in the next section. First, however, I wish to comment briefly on Gallagher's notion of body image as a form of primitive, pre-reflective awareness.

The proprioceptive awareness we witness in neonates can be considered to be the first manifestation of what we call the *mind*. However, Gallagher shows that it is nothing like the isolated, bodiless and static spectator that is usually presupposed by TT or ST. On the contrary, the mind as proprioceptive awareness, as a primitive body image, is structured and shaped by the body and its movement. It emerges as the result of *perception in action* - not in isolation, but through a continuous process of interaction with other minds. From the very moment of its conception, the mind can be seen as the expression of a self-consciousness that is at the same time already a *co-consciousness* (see fig. 4.4). Therefore, 'experientially, and not just objectively, we are born into a world of others' (Gallagher and Meltzoff 1996, p.226). ST and TT often argue that we need inferential and conceptual abilities to read the other mind, assuming that this is a prerequisite for intersubjectivity. But Gallagher shows that right from the moment of birth, children are *already* interacting with other minds. These interactions shape their minds in various ways, and provide them with a solid basis for future participation in more advanced social practices.

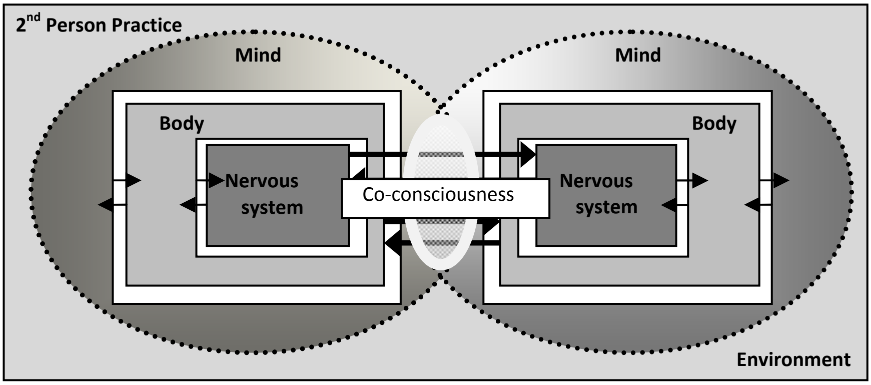


Fig. 4.4 Minds are already co-conscious from the moment of birth. Co-consciousness operates in between the semi-permeable bounds of embodied minds

4.2 Motor models and direct resonance systems

Motor models for basic adaptive feedback control

The challenge is to give a more detailed explanation of the relation between proprioception, body schema and body image, and demonstrate how they are embodied. In this section, I show how *functional motor models* can point us in the right direction.

Let us start by considering the very minimal and primitive body image that was introduced in the previous section. Gallagher (2000) calls this a 'minimal self', which he defines as a basic 'consciousness of oneself as an immediate subject of experience' (p.15).[65] He argues that the minimal self encapsulates two modalities of experience: (i) a sense of *ownership* (SO), the sense that I am the one who is undergoing an experience, and (ii) a sense of *agency* (SA), the sense that I am the one who is the initiator or source of the action.[66] How can we explain the relation between such a primitive body image and the body schema?

First, we need to know something about the motor theory of intentional action. This theory attempts to capture the dynamics of intentional action in terms of 'inverse', 'sensory-feedback' and 'forward' models (Blakemore and Decety 2001, Blakemore et al. 2001, Wolpert et al. 2001). The *inverse model* is important for motor control (see fig. 4.5). It consists of a simple sequence of steps, according to which a so-called 'planner' selects the appropriate motor commands given a desired goal (in terms of sensory states).

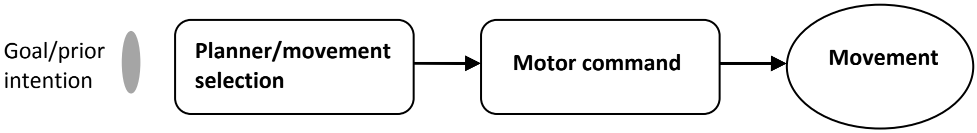


Fig. 4.5 Inverse model

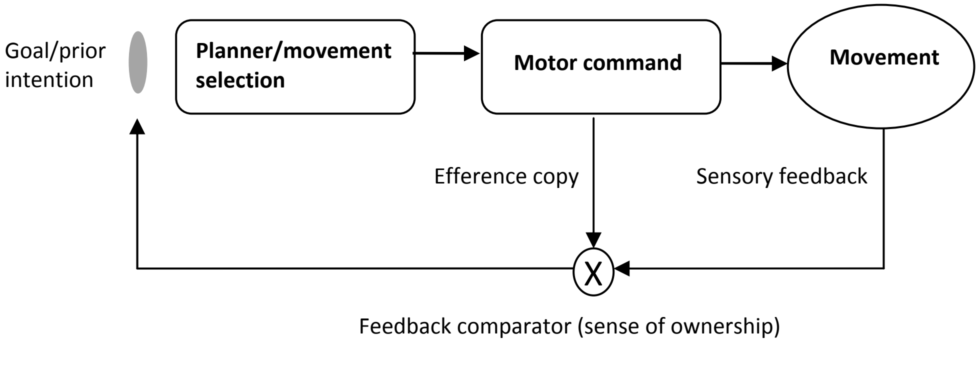
[65] Gallagher seems to have borrowed the term minimal self from Strawson (1999).

[66] In normal voluntary or willed action, SO and SA are intimately intertwined and often indistinguishable. However, Gallagher (2000) argues that there are a number of situations in which it becomes possible to distinguish between them, namely in cases of involuntary movements, unbidden thoughts, schizophrenic experiences such as thought insertion. In these cases, according to Gallagher, the sense of agency is lacking but the sense of ownership is retained in some form.

This motor command is then sent to the muscles, and this leads to movement.

The sensory-feedback model (see fig. 4.6) is an extension of the inverse model, because it contains an extra flow of proprioceptive information. When a motor command is sent to the muscles, an *efference copy* of this signal is sent to a self-monitoring system (or *comparator*), which compares it to *re-afferent sensory feedback* about the movement actually made. Feedback might include visual and proprioceptive inputs resulting from movements of one's own hands, or movement through space, or manipulation of objects. When there is indeed a *match* between efference copy and sensory feedback, the feedback comparator model delivers a sense of ownership (SO) for the action. Gallagher (2005) explains this as follows: 'Exteroceptive sense modalities (such as touch or vision) provide information about both the environment and the moving subject (tactile and visual proprioception). Such information comes into a complex intermodal relationship with somatic proprioception to form coordinated and intermodal sensory feedback. That sensory feedback coordinates with efferent copies of motor commands in the nervous system, verifying that it is the subject who is moving rather than the environment' (p.106). In this way, the sensory-feedback model is able to generate a 'non-observational and pre-reflective differentiation between self- and non-self' (p.175).

Fig. 4.6 Sensory-feedback model

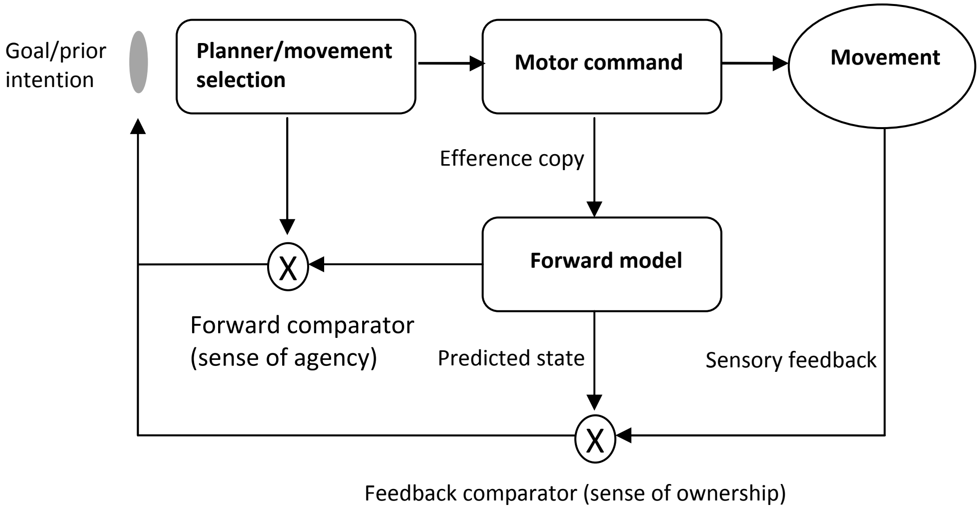


The sensory-feedback model is adaptive because it allows us to adjust ourselves to changing environmental conditions and compensates for exogenous disturbances: in the presence of different exogenous events, different outputs are needed to achieve the target.

This makes it possible to explain, for example, how we are able to correct our movement on the basis of sensory feedback about our actual movement. The model also sheds new light on the neonate ability to monitor and correct their imitations, in the sense that it shows that there need not be an *explicitly* recognized (cognitive) match between the infant's visual perception of the other's face and the proprioceptive awareness of its own face.

The sensory-feedback model is important for motor control, and explains how we are able to adjust our movements on the basis of sensory feedback. However, such an adjustment can only take place after the delays associated with sensory transmission. The so-called *forward model* bypasses these delays (and thus allows for better movement control) by positing a motor program that runs a slightly different sequence (see fig. 4.7). This time, the efference copy of the motor command is also sent to a *forward comparator*, which compares it to motor intentions and, when necessary, makes automatic corrections to movement *prior* to sensory feedback. Over time an association is established between efference copy and subsequent input, so that in effect a copy of the motor output signals comes to evoke the associated input signal.

Fig. 4.7 Forward model of goal-directed action


Goal/prior intention
Planner/movement selection
Motor command
Movement
Efference copy
Forward model
X
Forward comparator (sense of agency)
Predicted state
Sensory feedback
X
Feedback comparator (sense of ownership)

It can then operate as a *simulation* of feedback, to predict the consequences of output on input.[67] For example, the forward model might enable me to predict the sensory consequences of my act of reaching for and grasping a glass of water.

The forward model is responsible for generating a conscious sense of agency for action (Georgieff and Jeannerod 1998, Jeannerod 1994). Moreover, it can also account for the attenuated experience of the sensory consequences of one's own actions, compared to the sensory experience of exogenous changes in one's environment: the sensory consequences of one's own actions are predictable (from the efference copy of one's motor instructions), and therefore worth less perceptual attention than sensory changes exogenously produced. This could explain why in normal situations, proprioceptive awareness (PA) is attentively recessive and does not take center-stage in consciousness. This is because the forward model continues to function on the basis of proprioceptive information (PI), allowing one's body to work in a quite automatic way that does not require explicit monitoring. This may be different, however, in situations that require attentiveness to bodily movement. In infancy, for example, proprioception may be more centrally attended to when children learn how to walk. Early imitation also requires more focused propriospecific awareness. Gallagher and Meltzoff (1996) suggest that newborns use the proprioceptive experience of their own invisible movements to copy the movements of others. It not only helps them to monitor, correct and improve their imitations on the fly, but also allows them to memorize these imitations.[68] This shows that the forward model is not only important for motor control but also for motor *learning*.

The above models provide us with a functional architecture of the body schema, and they make it possible to explain how proprioception enables neonates to develop a primitive body image. But how is this architecture implemented at the neurobiological level? What kind of processes could facilitate the social interactions mentioned in the previous sections? And how do they provide young infants with co-consciousness, i.e. with an awareness of both self and other?

[67] Remark that this makes it possible to defend a (very weak) notion of simulation at the functional level. But such a notion depends on the intelligibility of forward models of goal-directed action, and it certainly does not satisfy a definition of simulation as the manipulation of pretend mental states.

[68] According to Gallagher and Meltzoff (1996), infants 'have the capacity to act out what they see in the face of the adult - they recognize what they see as one of their own capabilities' (p.223).

A neural architecture for imitation?

This is where the discovery of mirror neurons could be relevant. Mirror neurons are a class of visuomotor neurons that show activity in relation to both specific actions performed by self and matching actions performed by others. They 'mirror' the behavior of the other, as though the observer *himself* were acting (e.g., Rizzolatti et al. 1996, 2000). Mirror neurons appear to be involved in a larger cortical system (a 'mirror neuron system') that automatically 'duplicates' the observed action in the observer's motor system. This allows for an immediate, automatic and almost reflex-like understanding of others, without further inferential or conceptual requirements. Gallese (2001) gives the following explanation: 'when we observe goal-related behaviors […] specific sectors of our premotor cortex become active. These cortical sectors are those same sectors that are active when we actually perform the same actions. In other words, when we observe actions performed by other individuals our motor system 'resonates' along with that of the observed agent [...] action understanding heavily relies on a neural mechanism that matches, in the same neuronal substrate, the observed behavior with the one [the observer could execute]' (pp.38-9). What is attractive about the mirror neuron system is that it might explain how perception and action are dynamically co-constituted, and how action understanding emerges from the space that perception and action *share*.

Evidence for the existence of a mirror neuron system (MNS) was first discovered in the brain of the macaque monkey, comprising three cortical regions that exhibited the required functional properties and connectivity patterns: the superior temporal sulcus (STS) in the superior temporal cortex, area F5 in the inferior frontal cortex, and area PF in the posterior parietal cortex (Keysers and Perrett 2004).[69]

[69] In the early nineties, Perrett et al. (1989) demonstrated that neurons in the superior temporal sulcus (STS), which normally respond to moving biological stimuli (such as hands, faces and bodies) respond to these stimuli only when they are engaged in *goal-oriented actions* (see also Perrett et al. 1990, Perrett and Emery 1994). For example, some of them fired when the macaque saw a hand reaching toward an object and grasping it, but did not do so when the hand merely reached toward the object, without trying to grasp it. The investigators concluded from these observations that STS neurons probably code the perception of a meaningful interaction between an object and an intentional agent. The properties of these STS neurons seemed to be limited to the visual domain, since there was no association between the neuronal responses in STS and motor behavior. Another line of research, however, initiated by Rizzolatti et al. (2001), found parietal neurons with visual responses similar to the ones observed in STS *but this time with motor properties* (di Pellegrino et al. 1992, Gallese et al. 1996). These neurons were located in the ventral

Early experiments designed to detect the existence of a human MNS were motivated by the idea that if such a system existed, then the motor area it encompassed had to be active during both the execution and observation of a goal-directed grasping task. However, experimenters soon realized that instead of monkeyish grasping, *human imitation* offered a much more promising paradigm (Grafton et al. 1996, Rizzolatti et al. 1996). Imitation involves both the observation and execution of an action, and thus fitted perfectly with the properties of the system they were looking for. The investigators hypothesized that instances of imitation would yield an amount of mirror neuron activity approximately equal to the sum of activity during observation and execution. If they could identify brain areas that showed such a double amount of activity, then this would support the existence of a MNS in humans.

Iacoboni et al. (1999) found two areas that satisfied this condition, and also seemed to correspond anatomically to the macaque mirror areas. The first was located in the pars opercularis of the inferior frontal gyrus (in the inferior frontal cortex), the second in the posterior parietal cortex.[70] Together with the superior temporal sulcus (STS), coding for the perception of an observed intentional action, these areas could form a *blueprint* for the mirror neuron system. The STS, however, showed a somewhat unexpected pattern of activity. Although it yielded greater activity for action observation compared to control visual tasks and for imitation compared to control motor tasks (as was to be expected), there was also greater activity for imitation compared to action observation.

premotor cortex of the monkey (called area ‘F5’). Rizzolatti et al. (2001) also found that the posterior parietal cortex (PPC) of the macaque (area ‘PF’) contained mirror neurons almost identical to the ones described in F5. The areas PF and F5 appeared to be anatomically connected (Rizzolatti et al. 1998). Furthermore, evidence was found for a link between the STS neurons and the posterior parietal cortex (Seltzer and Pandya 1994). Together, the three cortical regions of the macaque brain (STS in the superior temporal cortex, area F5 in the inferior frontal cortex, and area PF in the posterior parietal cortex) seemed to have the functional properties and connectivity patterns required to instantiate a whole circuit for action recognition - a mirror neuron system.

[70] Iacoboni proposed a division of labor between the frontal and the posterior parietal mirror areas, inspired by single-cell studies (Kalaska et al. 1983, Lacquaniti et al. 1995) and neuroimaging data (Decety et al. 1997, Grèzes et al. 1998): the frontal mirror areas code the goal of the imitated action and the posterior parietal mirror areas code the associated movements. He claimed that certain experiments provide evidence for this idea. Koski et al. (2002), for example, demonstrated a modulation of activity in inferior frontal mirror areas during imitation of goal-oriented action, with greater activity during goal-oriented imitation compared to non goal-oriented imitation. See also the next section for a discussion of other experiments that might support such a proposal.

Mapping the above neural circuit onto the functional inverse and forward motor models described in the previous section allows us to make sense of the functional processes that underlie imitation (see fig. 4.8), and also helps us to understand the mentioned unexpected STS activity. Let us start with an observer who perceives the action of another agent. First, so-called *canonical* neurons in the superior temporal sulcus code an early visual 'description' of the perceived action (Perrett et al. 1990) and send this information to posterior parietal mirror neurons. This privileged flow of information is supported by robust anatomical connections between superior temporal and posterior parietal cortex (Seltzer and Pandya 1994). Second, the posterior parietal cortex codes the precise kinesthetic aspect of the movement of the agent (Kalaska et al. 1983, Lacquaniti et al. 1995) and sends this information to inferior frontal mirror neurons.[71]

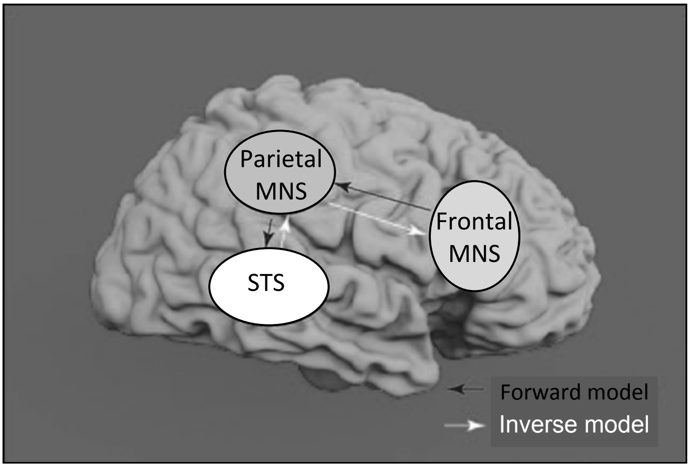


Fig. 4.8 A functional model of imitation (Iacoboni 2005):

1) The STS provides a higher-order visual 'description' of the observed action (inverse model)
2) This description is fed into the fronto-parietal mirror neuron system, where the goal of the action and the motor specifications to achieve it is coded (inverse model)
3) Copies of the motor imitative plan are sent from the fronto-parietal mirror neuron system to the STS, where there is a match between the predicted sensory consequences of the planned imitative action and the visual description of the observed action (forward model)

[71] Anatomical connections between these two regions are well documented in macaque monkeys (Sakata et al. 1973).

Third, the inferior frontal cortex of the observer codes the *goal* of the action. There is some neurophysiological (Umilta et. 2001, Kohler et al. 2002, Keysers et al. 2003) and imaging data (Koski et al. 2002) in support of this role for inferior frontal mirror neurons. This three-step process can be captured by means of a *forward* model, which uses the STS visual description of the action as input and the goal of the action as output.

Fourth, efferent copies of motor plans are sent from the parietal and frontal mirror areas of the observer back to the superior temporal cortex (Iacoboni et al. 2001), such that a matching mechanism between the visual description of the observed action and the predicted sensory consequences of the planned imitative action can occur. And fifth, if there is a positive match between the visual description of the observed action and the predicted sensory consequences of the planned imitative action, this forward/inverse model is reinforced by a 'responsibility signal' (Haruno et al. 2001) that assigns high responsibility for imitating the desired action. The observer is now ready to imitate the action of the other agent.

The mirror neuron system and action understanding

The above blueprint of the MNS might help us to give a plausible explanation of infant imitation. But we need to be careful here. To start with, there are different ways to make sense of imitation. The most restrictive definition of imitation requires the execution of a *novel* action (that is learned by observing another do it) and, in addition to novelty, also involves some understanding of the *means/ends structure* of that action: you have to be able to copy the other's means of achieving her goal, not just her goal, or just her movements. Research on human imitation has shown that infants of 13-14 months are able to do this. But although the human MNS resembles the system found in the macaque brain, macaque monkeys are not able to imitate in the strict sense. They only have the capacity for action *emulation*. In action emulation, you observe another person achieving a goal in a certain way, find that goal attractive and attempt to achieve it yourself by whatever means.[72]

[72] Note that the reproduction of an observed action may be the same whether it is performed by imitation or emulation (cf. Czibra 2007). If the observer has effectors and biological constraints similar to that of the model, it is likely that she will emulate the outcome of the model's action by

One important question is whether the human MNS by itself is able to facilitate imitative behavior in the strict sense of the word. As Hurley (2008) points out, Iacoboni's model is *in theory* able to explain how we understand the means/end structure of an action, because it distinguishes between the neural coding of *goals* and *movements.* When an observer perceives an agent moving in a goal-directed way, his inferior frontal mirror neurons encode the *goal* of the observed action, and this provides him with an understanding of the *intention* of the action. In addition, his posterior parietal mirror neurons encode the *movements* associated with the observed action, and this provides him with an understanding of *how* to achieve the goal by means of the observed movements. Linking these two processes in the right way could pave the way for imitative learning (in the strict sense).

However, in practice it turns out that, besides the MNS, imitative learning involves other brain areas as well. Molnar-Szakacs et al. (2005), for example, found that the imitation of novel actions yields additional activation of the dorsolateral prefrontal cortex (BA46) and cortical areas that are involved in motor preparation: the dorsal premotor cortex, the mesial frontal cortex and the superior parietal lobule. They argued that the activity in BA46 seemed to reflect the selection of motor acts that are 'appropriate' for the task that is executed (cf. Rowe et al. 2000).

In order to find out to which extent the MNS supports the understanding of action *intentions*, we have to explicate the notion of intention first. Gallese and Goldman (1998) originally proposed that the MNS might explain intentional action understanding in terms of *propositional attitudes.* They hypothesized that in case of a plan 'externally generated' in the brain of the observer, the latter's mirror neuron system would *retrodict* the 'target's mental state' (i.e., the agent's intention) by 'moving backwards from the observed action' (pp.495-6). However, the ability to understand actions in terms of propositional attitudes is a rather advanced mode of social interaction. Hutto (2009) remarks that it is a 'sophisticated high level capacity; it involves being able to answer a particular sort of 'why'-

means of the same behavior, i.e. she will faithfully reconstruct the observed action. This is why, in studies of imitation, unusual or inefficient goal-directed actions are demonstrated to participants in order to test whether they tend to *emulate* the outcome by their own (more efficient) means, or really *imitate* the observed action (Meltzoff 1988, Gergely et al. 2002, Horner and Whiten 2005). Meltzoff (1988), for example, tested whether infants are capable of imitation by demonstrating an unusual action to them, in which the model switched on a box-light by pushing it with his forehead. If the infants emulated the outcome, then they would have used a simpler action to achieve the same goal, such as pushing the box with their hands (cf. section 4 of this chapter).

question by skillfully deploying the idiom of mental predicates (beliefs, desires, hopes, fears, etc.)' (p.10).

Nevertheless, Iacoboni et al. (2005) have tried to demonstrate that the MNS contributes to the understanding of the 'why' of an action as well. In their experiment, they presented subjects with a series of short movies, which were labeled the 'context' condition, the 'action' condition and the 'intention' condition. In the context condition, subjects would see objects (a tea-pot, a mug, cookies, etc.) arranged either as if before tea (the 'drinking' context) or as if after tea (the 'cleaning' context). In the action condition, subjects would see a human hand grasp a mug either with precision grip or using a whole-hand prehension with no other contextual elements present. In the intention condition, the grasping actions were embedded in the two scenes used in the context condition, the drinking context and the cleaning context. Here, the context cued the intention behind the action.[73]

Iacoboni et al. (2005) found significant differences between the intention condition and the action and context conditions in the human brain areas known to have mirror properties. They showed that, compared to the action condition, the intention condition yielded significant signal increases in visual areas (STS) and the dorsal part of the pars opercularis of the inferior frontal gyrus. Importantly, they also found increased activity in the pars opercularis.

The experimenters argued that this means that Iacoboni's model is basically correct in assuming that the MNS does not simply enable movement recognition ('that's a grasp'), but also is critical for understanding of the *goal* of an action. According to them, the experiment shows that the MNS not only enables the observer to understand *what* the agent is doing (by generating motor activation associated with the same *movement* in the observer), but also *why* the agent is doing this (by generating motor activation associated with a similar *goal* for the observer).[74] They conclude that 'the role of the mirror neuron system in coding actions is more complex than previously shown and extends from action recognition to the coding of intentions' (p. 532).

[73] The 'drinking' context suggested that the hand was grasping the cup to drink. The 'cleaning' context suggested that the hand was grasping the cup to clean up. Thus, the intention condition contained information that allowed the understanding of intention, whereas the action and context conditions did not (since the action condition was ambiguous, and the context condition did not contain any action).

[74] Cf. Iacoboni et al. (2005), see also Rizzolatti and Craighero (2004).

These findings seem to indicate that the MNS plays an important role in the understanding of intentional action. However, we have to remark that the above experiment still deals with *intentional* instead of *propositional* attitudes. In this respect, Iacoboni et al. seem to be interested in a notion of intentionality that is far more basic than the one that Gallese and Goldman (1998) were originally after. There are other problems as well. According to Jacob (2005), for example, it is possible that the enhanced activity found in the pars opercularis of the observer is not so much the *output* but the *input* of the mirroring process. Csibra (2007) has pointed out that this problem reveals a more general tension between two conflicting claims made by those who defend an all-inclusive approach to intentional action understanding in terms of mirror neurons: namely, that action mirroring somehow is thought to represent *both* low-level resonance mechanisms *and* high-level action understanding. This tension arises from the fact that 'the more it seems that mirroring is nothing else but faithful duplication of observed actions, the less evidence it provides for action understanding; and the more mirroring represents high-level interpretation of the observed actions, the less evidence it provides that this interpretation is generated by low-level motor duplication' (p.447).[75]

I am not sure whether Csibra's criticism hits home. Actually, I think that his ideas about what it means to understand the intention behind an action are quite different from those of Iacoboni et al. - perhaps they come closer to Hutto's notion of propositional attitudes. But Csibra is probably right that we need more than just the MNS in order to

[75] Csibra (2007) illustrates this point by discussing another experiment on intention understanding in monkeys by Fogassi et al. (2005). In this study, single cell recordings were used in an attempt to prove that the MNS enables the observer (a monkey) to discriminate between two tokens of an intentional act of grasping, one for the purpose of eating and another for the purpose of placing. Csibra (2007) points out that the slight kinematic variation that was found (though not reported) could explain the activation difference across the mirror neurons. If we assume that the observed actions included the same kinematic differences as in the monkeys' actions, and the monkeys' parietal mirror neurons were sensitive to these parameters, then their activation represents a low-level mirroring phenomenon. According to Csibra, however, in this case nothing suggests that the monkeys understood the *intention* behind the observed action. But if we accept that the selectivity of the mirror neurons was independent of the kinematic parameters and reflected a true form of intentional understanding based on contextual cues, as Fogassi et al. (2005) would have it, then there is no evidence that this intentional understanding is based on low-level mirroring. Therefore, Csibra (2007) concludes that 'one cannot have one's cake and eat it too: the discharge of a set of MNs cannot represent the activation of the observer's motor system at low and high levels at the same time' (p.447).

account for findings on intentional action understanding such as those of Iacoboni et al. (2005).

Originally, mirror neurons were simply supposed to fire during both the execution and the observation of one and the same motor act, and the resulting match was thought to be responsible for action understanding. But many MNS advocates now recognize that if Iacoboni's model of action understanding is correct, then we need to explain the relationship between the neural coding of intention and movement *given a certain context*. According to Iacoboni et al. (2005), this requires an 'additional mechanism', one that involves neurons that are 'perceptually triggered by a given motor act', but whose discharge commands the execution of a different motor act 'functionally related to the former and […] part of the same action chain' (p.533). They suggest that the coding of action intentions is probably 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, given a certain context. However, Jacob (2005) correctly points out that the idea of a motor chain that consists of 'logically related' mirror neurons sounds pretty much like a classical *inference* system that translates between perception and action. Such a mechanism only seems to be required because the MNS by itself cannot account for the complexity of mapping an observed movement onto an underlying intention.

In other words: it is not yet clear to which extent the MNS is involved in the context-sensitive understanding of goal-directed action, and this is subject to further research. However, a somewhat simplified version of Iacoboni's blueprint could still be used in order to explain how we are able to *anticipate* the actions of others. If we use a two-step inverse-forward model, it is possible to explain action anticipation in the following way: first, the STS feeds a visual description of the perceived action into the fronto-parietal mirror neuron system, where the kinaesthetic aspect of the perceived movement is coded in terms of a motor plan (inverse model). Second, efferent copies of this motor plan are sent back to the STS in order to predict the sensory consequences of this action (forward model). This process allows the observer to predict and anticipate the agent's next move, i.e., his next motor sequence.

Interestingly, the idea that the MNS might enable action *anticipation* in accordance with this simplified inverse-forward model perfectly fits with Gallese and Goldman's (1998) original suggestion that the ability to anticipate and predict the next move of conspecifics is crucial, since this move might be 'cooperative, non-cooperative, or even threatening' (pp.495-6). However, my proposal requires a much weaker (non-conceptual) notion of intentionality, and therefore offers a much more pragmatic explanation of the kind of intentional action understanding that young children display.[76]

Some additional considerations

The ability to anticipate the other's movement can be regarded as enabling a very early, action-oriented stage in the development of our understanding of other minds. But the MNS could also help us to explain our occasional feelings of *empathy* - the 'subjective experience of similarity between the self and others' (Decety and Jackson 2004, p.71).When we perceive the emotions and sensations of other agents, our MNS might trigger our motor system to resonate along with that of the observed agent in a direct fashion, and make it possible to 'put ourselves in the other's shoes' in order to experience what they are feeling in a non-inferential and non-conceptual way.

There is evidence that this is indeed what happens. Wicker et al. (2003), for example, showed that mirror processes play an important role in our experiencing of the emotion of disgust. They scanned participants both during their own experiences of disgust and during observation of other people's faces expressing disgust. The participants were scanned while viewing movies of individuals smelling the contents of a glass (disgusting, pleasant, or neutral) and forming spontaneous facial expressions. The same participants were also scanned while inhaling disgusting or pleasant odorants through a mask. The experimenters found that the same areas, the left anterior insula and the right anterior cingulate cortex were preferentially activated both during the experience evoked by disgusting odorants and during observation of other people's disgust-expressive faces. Gallese (2001, 2004) found similar evidence for the relevance of mirroring processes with respect to the sensation of pain.

I wish to close this section with a final observation about the importance of imitation and the involvement of the MNS. Full-fledged imitation allows young children to copy both

[76] Remark that the fact that action anticipation can be explained in terms of an inverse/forward model does not provide any support for ST, since it is not possible to draw a strict line between the observation of an action and something that counts as a simulation. Nor does this kind of action anticipation involve any mental state management. See also chapter 2.4.

the means and the goal of the actions they observe, and this is arguably a great way to generate new possibilities for intersubjective understanding without having to appeal to an innate theory of mind. In his discussion of Fodor's nativism, Meltzoff (2005) remarks that: 'Fodor is correct that solipsism and blank-slate empiricism are too impoverished to characterize the human starting state. However, this does not mean that adult commonsense psychology is implanted in the mind at birth or matures independent of experience. Here is an alternative to Fodor's creation myth. Nature designed a baby with an imitative brain; culture immerses the child in social play with psychological agents perceived to be "like me". Adult commonsense psychology is the product' (p.77).

At the same time, however, we have seen that we should not expect an easy explanation as to precisely what it involves. Although the MNS might be a step in the right direction, there are certainly more aspects that need attention. One of them concerns the role of *inhibition* in the imitation of behavior. In Iacoboni's model, for example, there needs to be a positive match between the visual description of the observed action and the predicted sensory consequences of the planned imitative action, otherwise the perceived action will not be imitated. But this already presupposes a form of inhibition that is quite advanced. How are these inhibitory processes related to the MNS? This requires further research. Another important aspect has to do with the experimental set-ups that are used to investigate the role of the MNS in our intersubjective engagements. Currently, most experiments on the MNS involve the passive third-person observation of another agent. But this is clearly not representative for most of our everyday social encounters, which are better characterized in terms of active second-person engagements. Future research on the MNS definitely has to take this into account.

4.3 Embedded practices

Shared attention

The previous sections demonstrated that neonates and very young infants are well able to individuate other persons and interact with them in a *dyadic* way. But arguably these 'primary intersubjective' capacities (Trevarthen 1979) do not yet involve a very strong notion of intentional understanding on behalf of the young infant. It is generally accepted that such a notion only starts to emerge when infants start perceiving and interacting with other agents in a world-involving way, entering the realm of 'secondary intersubjectivity' (ibid.). The embedded practices that are characteristic for secondary intersubjectivity are *triadic*, in the sense that they involve a referential triangle of child, adult, and the environment - an outside object or event to which they share attention (see fig. 4.9).

Shared attention not only involves infants attending to the same objects or event at the same time, however. It also requires that they mutually recognize that their attending has a *common focus*. This makes it quite different from the forms of reciprocal imitation described in the previous sections. In situations of shared attention, according to Hutto (2007a), 'I see what the other is attending to, I see that they are attending to it, and I see that they are attending to both the object and to my attending. Only in this way is the object recognized as a common focal point' (p.126).

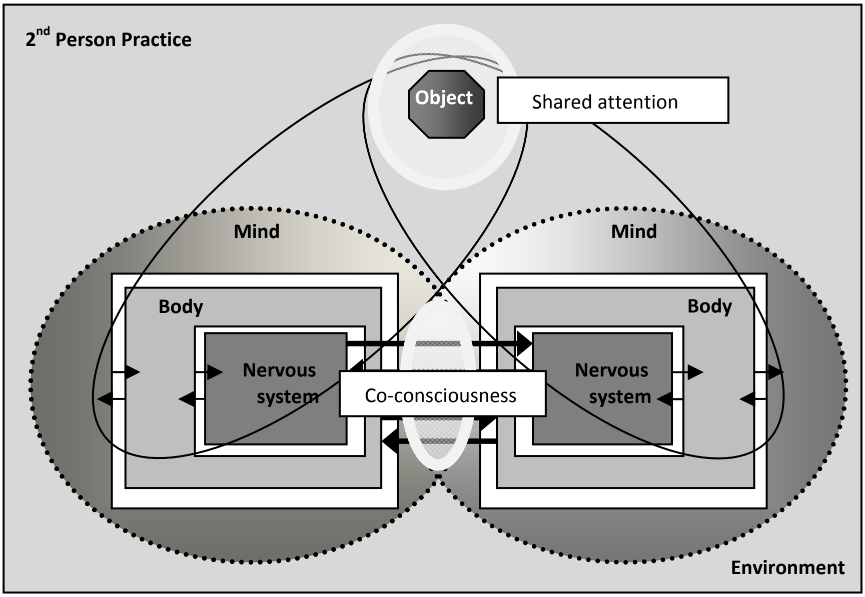


Fig. 4.9 Shared attention in secondary intersubjectivity

From 6 months of age onwards, infants are already capable of perceiving other people as being directed towards objects, first in their grasps of objects, later also when they gaze and point at (distant) objects (Woodward 2005). Yet, as Tomasello et al. (2005) point out, infants' object-directed understanding of others merely has the effect that they 'expect the adult to be consistent in his interactions with the same object over a short span of time [...] they do not have any understanding of the internal structure of intentional actions' (pp.678-9). But by 12 months of age, experimental findings suggest that infants 'can (1) interpret others' actions as goal-directed, (2) evaluate which one of the alternative actions available within the constraints of the situation is the most efficient means to the goal, and (3) expect the agent to perform the most efficient means available' (Gergely and Csibra 2003, p.288). However, Reddy (2003) has argued that shared attention can be said to arrive much *earlier* as long as it is not defined with respect to an outside object but rather to the child itself (since the child is already aware that itself can be an object of attention). Interestingly, she suggests that we have to pay more attention to the *second-person interaction*s between the infant and the experimenter to discover this.

Baron-Cohen (1995) has proposed that our capacity for shared attention might be facilitated by a 'shared attention mechanism (SAM)'. He argues that 'SAM's key function is to build […] triadic representations. Essentially triadic representations specify the relations among an Agent, the Self, and a (third) Object. [...] Included in a triadic representation is an embedded element which specifies that Agent and Self are both attending to the same object' (pp.44-5).

Again, the question is why we would need *mental representations* to explain what happens in these cases. As Hutto (2006) argues: 'There is no reason to suppose that [...] shared attention involves making full-fledged propositional attitude ascriptions. Seeing another's seeing does not involve representing the other's *cognitive* take, it only requires recreatively imaging the other's *perceptual* one' (p.192, italics in original). What underpins the mutual connectedness that is characteristic for shared attention is probably much better explained in terms of the MNS.

Pre-linguistic behavior representative of shared attention includes the systematic use of communicative gestures for instrumental purposes such as pointing and gaze alteration (Butterworth and Grover 1990, Butterworth and Jarrett 1991). Infants not only flexibly and reliably look where adults are looking (gaze following), but also try to obtain emotion cues from others to assist in their own assessment of an uncertain or ambiguous situation – this is called ‘social referencing’ (e.g., Rosen et al. 1992). Some have argued that the latter ability requires a ‘rudimentary ability to impute mental states of self and other’, and on top of that, a basic understanding that ‘one mind can be interfaced with another’ (Bretherton 1991, p.57). However, it is very well possible to give an explanation of social referencing without assuming that the infant has to attribute mental states to others. For example, observing the mother's emotion expression may induce the corresponding emotion in the infant who can then proceed to appraise the ambiguous situation on the basis of its *own* felt emotion. What is problematic about the appeal to infant's instrumental use of others to achieve its goals as evidence for attributing mental states to persons is this: while this kind of behavior indeed indicates that the infant is an intentional agent, it clearly does not imply that it must perceive the other person as an agent whose actions are caused by intentional mental states that the infant manipulates through its communicative gestures.

Basically, what happens in shared attention is that children’s ability to perceive affordances, i.e. to see objects in the environment as inviting or as enabling certain kinds of actions, is re-centered to the perspective of the other. Instead of seeing things as affording something for *themselves*, they now see them as affording something for *others* as well. This lays a foundation for the more advanced modes of perspective taking that are characteristic for narrative practice.

Further developments

Perceiving the other as an intentional agent who is responsive to a growing array of affordances in their environment allows the infant to anticipate, cooperate and coordinate in increasingly complex practices. In situations of shared attention, several behaviors often come together, enabling infants to ‘tune in’ to the attention and behavior of adults toward outside objects and events (cf. Tomasello 1999). Amongst others, there is a further development of imitative behavior - infants begin to act on objects in the way adults are acting on them. Already at 6 months of age infants can reproduce others’ actions on objects (Barr et al. 1996). However, it is not until the age of 13-14 months that there is evidence of *imitative learning*. Imitative learning consists of reproducing the intentional actions of others, including both the goal at which they are aiming and the behavior or strategy by means of which they are attempting to accomplish that goal. For example, in a study by Meltzoff (1988), infants of 14 months old observed an adult bend at the waist and touch his head to a panel, thus turning on a light. The infants followed suit even though they might also have turned on the light by simpler means (e.g., with their hands) - implying that they were indeed reproducing the adult's action. Moreover, they did not turn the light on in this odd way unless they had seen the model do it first (see also Meltzoff 2004, Gergely et al. 2002). Similarly, 16-months-old infants imitatively learn from a complex behavioral sequence only those behaviors that appear intentional, ignoring those that appear accidental (Carpenter et al. 1998). They do not just mimic the limb movements of other persons; they attempt to reproduce other persons' intentional action. By 18 months, infants are able to re-enact to completion the goal-directed behavior that an observed subject does not complete (e.g. pulling apart miniature dumbbells), but they will not re-enact the target act when it is performed by a mechanical device (Meltzoff 1995, Meltzoff and Brooks 2001).[77]

Also emerging in the second year is the infant's capacity for *pretend play.* The main characteristic of pretend play is that children pretend an object to be something else (Leslie 1987, Garvey 1990, Lillard 2002). For example, a child who is pretending a pile of sand is fantastic chocolate cake might call it cake, mimic eating it, and perhaps even say: 'Yum-yum, what delicious cake!' Notably, the child will not actually go so far as to eat the sand, since it is clearly aware of the cake's real identity. Around this age, children also become capable of recognizing pretend behavior of *others*. For example, when the mother pretends the banana is a telephone, the child is able to pick it up, hold it up to his ear and mouth and say: 'Hi. How are you? [Brief pause] I'm fine. OK. Bye'. Besides the substitution of objects, pretend play can also involve imagined objects, or roles and situations.

It has been argued that pretend play presupposes a capacity for 'secondary representation'. Perner (1991), for example, claims that young infants are initially only capable of entertaining 'primary representations' that represent 'the world as it is'. Such a representational system is not sufficient to facilitate pretend play, however. The child's primary representation of a banana, for example, cannot also incorporate a representation of this banana as a telephone. This requires the capability to entertain 'secondary

[77] Interestingly, infants are more prone to imitate an unfulfilled goal if the action is marked linguistically as purposeful, e.g. 'Let's put this on here. There we go!', but not if it is marked as accidental, e.g. 'Let's put this on here. Whoops!' (Carpenter et al. 1998). Infants not only try to reproduce the intentional *actions* of others, but also pieces of *language* (Tomasello et al. 1996).

representations', which add the ability to model hypothetical situations, and makes it possible for the child to simultaneously entertain multiple mental models.

However, Lillard (1993) found that children first understand pretending only as an action, and only much later come to see it as involving mental representations. In fact, most studies of pretense involve pretense with actions (cf. Flavell et al. 1987, Wellman and Woolley 1990). In these experiments, children perform correctly by directly referring to the *actions themselves*, rather than by *mentally representing* them. In other words, they initially seem to interpret pretense in terms of action alone. This emphasizes the embodied, situated and enactive character of early pretend play.

What is problematic about the notion of mental representation is precisely that it is usually associated with a number of *opposite* features, such as context-independency (representational contents are self-sufficient and exportable to different situations), objectivit*y* (representations depict in a isomorphic way how the world and the actions are structured) and abstractness (representations provide neutral depictions valid under any possible perspective, not from situated points of view).

Shared attention and language acquisition

Shared attention is an important precursor to the development of *linguistic practices*. Infants pick up language socially by using it in pragmatic context, and by noticing what others do with it, through sharing interests, pretend play and imitative learning (Bates et al. 1975, Ninio and Snow 1996). There is much evidence that shared attention is strongly associated with the picking up of words in the infant's second and third year (Locke 1993, Rollins and Snow 1998, Tomasello 1988). Its onset not only consistently precedes the emergence of referential language in the second year of life, but the ability to engage in shared attention during infant-mother interactions also predicts the infant's word comprehension and word production (Carpenter et al. 1998).[78]

New words also prepare children for more sophisticated social interaction. Eilan (2005) observes that the 'first words emerge during the thirteenth month, on average, and

[78] For example, gaze following, an important prerequisite for shared attention, predicts vocabulary between the first and the second year (Morales et al. 1998), and shared attention bids have been shown to make a unique contribution to language development at 30 months (Morales et al. 2000).

from then on until the end of the second year, attentional behaviors become progressively more sophisticated - for example, we find progressively sensitive checks of where the adult is looking, before, during and after pointing initiated by the infant, or showing of objects to adults, the bouts of attending together to an object become longer and able to sustain the beginning of extended play with, and conversations about the object(s) attended to' (p.5). The growing ability to use linguistic signs provides children with new modes of expression and enables more advanced forms of understanding others than those of the purely embodied and embedded form.

Although it is often assumed that young children acquire language through ostensive definition (adults stop what they are doing, hold up objects, and name these objects for them), this is empirically not the case. In general, for the vast majority of words in their language, children must find a way to learn them in the ongoing flow of social interaction, sometimes from speech not even addressed to them. Tomasello et al. (1993) call this kind of imitative learning *cultural learning* because the child is not just learning things *from* other persons; it is also learning things *through* them in the sense that it must know something of the adult's perspective on a situation to learn the active use of this same intentional act. The idea is that children only come to understand a symbolic convention by learning to understand their communicative partner as an *intentional agent*, one with whom one might share attention, since 'a linguistic symbol is nothing other than a marker for an intersubjectively shared understanding of a situation' (Tomasello 1999, p.516).

The development of linguistic practices does not only depend on the embodied and embedded practices described in the sections above, but it also takes them to the next level. Language starts to provide 'an immensely delicate and useful way of pointing' (Heal 2005), exponentially extending the ways in which infant and other can explore the world together, adding rationales of increasingly complex structure to the world of the infant, possible reasons that they and others may act upon. How this happens is the topic of the next chapter.

4.4 Social understanding without cognitive or conceptual requirements

Summary

In this chapter I have discussed a number of embodied and embedded practices through which infants learn to deal with others in a direct, i.e. non-conceptual and non-inferential fashion. These interactions contextualize our engagements with other minds and provide us with the 'know how' that is required for more advanced (conceptual) modes of intersubjectivity. Importantly, they are best and most parsimoniously explained without reference to theory, simulation or mindreading. I very much agree with Hutto (2007a), who stresses that our 'nonverbal acts of intersubjective responding are not prosecuted by the deployment of theory, inferential reasoning, or projective simulation. We can be sure of this because no ascriptions are made to others on the basis of their observed behavior - there is no need to bridge an imagined gap between self and other; indeed the very idea of such a gap existing at this level is problematic' (p.115). Embodied and embedded practices do not presuppose higher order cognitive abilities or advanced mental state manipulation skills. Rather, they structure and scaffold these later developments.[79]

I have shown that in their 'ordinary' second-person interactions with others, children do not put themselves in the observer position – they are not passively standing at the side thinking about how to access other minds or trying to find explanations for others' behavior. Rather, they actively respond to them in various embodied and embedded ways (see fig. 4.10). Gallagher (2007) hits the nail on the head when he claims that 'what we call social *cognition* is often nothing more than 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' (p.540, italics added). These perception-action loops are structured and shaped by our bodily existence and various (partly) innate sensory-motor capacities. Mirror neuron processes show how perceived

[79] Nor do these practices merely function as developmental precursors to a Theory of Mind. There is only one study that reports an association between pretend play in 33-months-old children and their success in passing a series of false-belief tasks 7 months later (Youngblade and Dunn 1995). Two other studies, however, fail to reveal a similar longitudinal association between pretense and Theory of Mind development (Charman et al. 2000, Jenkins and Astington 2000). With regard to imitation, there is no evidence for an association between early imitation and the later development of a Theory of Mind (Charman et al. 2000).

behavior and responsive actions can become intelligible *together*, that is, in the same process. They allow for a dynamic co-constitution of perception and action, and do not require inferential or conceptual process to mediate between them. At the same time, however, our intersubjective abilities cannot be explained in (or reduced to) purely neurobiological terms. Although brain processes are without a doubt important for explaining how infants are able to understand others, they would not occur unless these infants were acting within a broader social context. This context has to be taken into account in order to do justice to the interactive nature of intersubjectivity.

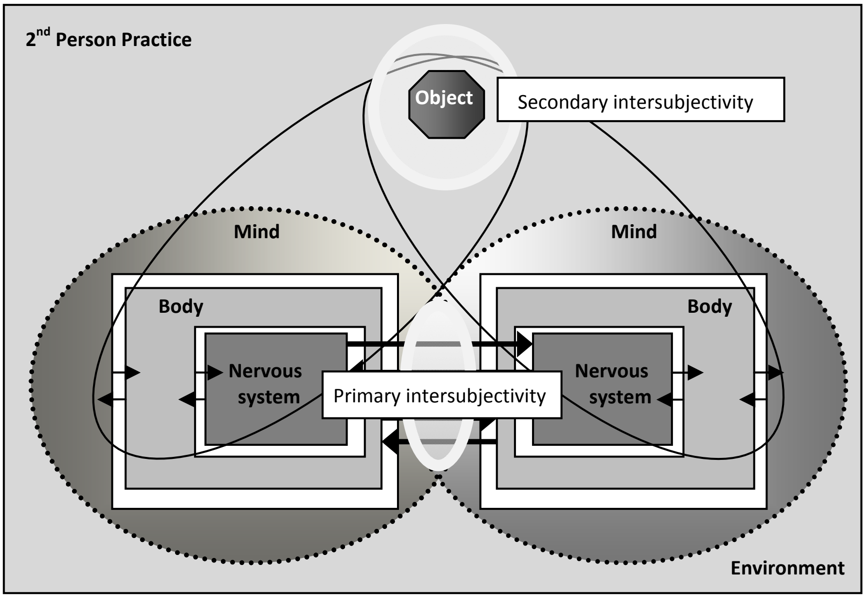


Fig. 4.10 Primary and secondary intersubjectivity

Primary intersubjectivity:

imitation
intentionality detection
action anticipation
eye tracking
movement tracking
emotion understanding

Secondary intersubjectivity:

shared attention
pointing / gaze alteration
social referencing
agency detection
pretend play
advanced imitation

Direct perception

In the remainder of this chapter, I wish to address the question whether we need to further articulate the embodied and embedded practices described in the previous sections, and if so, how. Gallagher (2001) has proposed the term 'body reading' in order to stress the perception-based nature of understanding that is characteristic for these practices. He claims that during our intersubjective engagements, it is very likely that 'various movements of the head, the mouth, the hands, and more general body movements are *perceived* as meaningful or goal-directed [...] such *perceptions* are important for a non-mentalistic (pre-theoretical) understanding of the intentions and dispositions of other persons [...] In *seeing* the actions and expressive movements of the other person one already *sees* their meaning; no inference to a hidden set of mental states (beliefs, desires, etc.) is necessary' (p.90, italics added).[80]

In his recent writings, Gallagher has further extended these ideas into a theory of 'direct perception' (cf. Gallagher 2007). His starting point is the observation that TT and ST approaches to intersubjectivity somehow seem to assume that perception is by itself not *sufficient* for social interaction. Something more is needed in order to understand our fellow human beings, and this is the reason why these positions appeal to mindreading procedures. According to Gallagher, the problem is that TT and ST start with a notion of perception as 'third-person observation', rather than something that happens in the context of *interaction*. As a result, we are not actively involved with others, but we stand at 'the margins of the situation.' However, Gallagher argues, this idea of perception as mere observation leaves TT and ST with an extremely impoverished idea of what perception actually consists in when it comes to perceiving other people. 'If I were to remain with only this perception I would be totally perplexed or at least puzzled about the other person's behavior. I see what the other person does, but until I call forth some theory, or until I run

[80] Hutto (2007) points out that the reading metaphor is misleading here, since it retains an 'intellectual connotation' that misrepresents what goes on in basic intersubjectivity and ignores that, although infants are not *reading* minds, they are 'immediately responsive to "other minds" nonetheless' (p. 116). I agree, but I think what is more problematic is the emphasis on *individual perception*.

through a simulation routine, I seem not to have any sense of what that person is up to' (p.536).[81]

Gallagher's own approach, by contrast, depends on a very rich notion of perception that builds on the idea that we have a direct perceptual grasp of the other's intentions, feelings, etc. The kind of perception Gallagher has in mind is 'direct' in the sense that nothing is added to it. When we see the actions and expressive movements of other persons, we are able to *directly* perceive their meaning. We do not need to consult a folk psychological theory or run a complicated simulation routine.

The question is why we have to place so much emphasis on the role of direct perception. Although I welcome Gallagher's rejection of mindreading when it comes to explicating low-level embodied practices, I think Hutto (2007a) has a point in claiming that it is 'more correct to say that we are directly *moved* by another's psychological situation rather than that we directly *perceive* it' (p. 116, italics added). The problem with the notion of direct perception is that, despite its success in overcoming the TT/ST mindreading legacy, it seems to encourage an interpretation of second-person embodied practices in terms of *individual perceptual* capacities. And this brings back the old idea that intersubjective understanding is primarily a one man (or woman) spectator sport, a social 'know-how' that is modeled on the first-person perspective of the individual agent.

Perception and the intrasubjective bias

Many of the problems that trouble TT and ST approaches to intersubjectivity can be traced to their commitment to a strong form of internalism, and the (Cartesian) ideal to model our understanding of others on the mind/brain of the individual agent. If we are to avoid these problems, we have to reject the idea that intersubjectivity is primarily a *personal* achievement, and maintain a clear focus on the second-person nature of social interaction. The challenge for a pragmatist approach to intersubjectivity is to look beyond the embodiment of individual subjects in order to properly conceive of the embedded and interactive nature of our understanding of others. This is important because it can put a

[81] Importantly, Gallagher recognizes that this kind of perception is not *completely* impoverished, since it is still smart enough to allow the agent to distinguish between an object in the surrounding environment and another agent. But this is not sufficient for the kind of social understanding that TT and ST are after: some 'extra cognitive tools' are required.

stop to simplistic ‘just so stories’ about intersubjectivity, and prevents all too easy ‘explanations’ of the acquisition of our social abilities.

What is problematic about the notion of direct perception is precisely that it seems to discourage us to look beyond individual embodiment. Gallagher’s (2000) ideas about the *minimal self* seem to confirm this worry. The word ‘minimal’ is usually employed to denote the most limited case we can come up with. Therefore, the notion of a minimal self seems to suggest that we can just slice up the self into small pieces in order to find its most minimal part. However, it is very probable that in the end this part will refer to an *individual agent*. The quest for a minimal self easily runs the risk of slicing off the *social* dimension, thereby leaving us with the primacy of an impoverished first-person perspective. This in turn enhances the idea that we need some kind of ‘self-sufficient self’ before we can engage in social interaction. To avoid these mistakes, it is probably much better to speak of a ‘basic self’ instead of a ‘minimal self’ – one that is thoroughly social and relational. Since our experience is always intentional and directed at something, it is necessarily *relational.* And since many of our interactions are with other persons, this relatedness is also a social relatedness. At its most basic stage, the self is always already ‘co-conscious’. Moreover, it always already finds itself (or is ‘thrown’, to use the Heideggerian terminology) in a social *practice.* My wariness of direct perception precisely stems from the conviction that our interactions with others cannot be explained in terms of capacities that are purely *individual or intrasubjective*. Of course I am not denying that there are such things as individual capacities, but the important question is how we acquire them.

Directness versus development

Whereas the term ‘perception’ seems to be unsuitable for an account of social interaction because it suggests the primacy of the first person perspective, the term ‘direct’ has the drawback of suggesting that social interaction is never *problematic*, since there is always an immediate and direct understanding of the other. For TT and ST accounts of intersubjectivity, our understanding of other minds is always indirect and deeply problematic. Consequently, we need a lot of theoretical back-up in order to survive our social encounters, and our success on this score is measured by our ability to predict others’ behavior. Gallagher, by contrast, seems to suggest that social sense-making is in principle easy and effortless from the very moment we are born. He argues that what is important about direct perception is not ‘what directness means, but how smart, how richly informed, it is. 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’ (p.538). Unsurprisingly, the kind of perception Gallagher wishes to promote is very smart and richly informed.[82] He claims that ‘practically speaking, direct perception, etc. delivers what I need to interact with others most of the time. In the broad range of normal circumstances there is already so much available in the person’s movements, gestures, facial expressions, and so on, as well as in the pragmatic or social context, that I can grasp everything I need for understanding in what is perceptually available’ (Gallagher 2007, p.540).

But how and where did direct perception become so incredibly smart? This problem appears to be a direct consequence of Gallagher’s intention to model intersubjective knowledge on first-person perception. Admittedly, Gallagher states that one of the sources of intelligent perception is social experience, which fine-tunes our sensory-motor neuronal systems. Also, he acknowledges that direct perception gains in intelligence as infants develop, acquire language, conceptual competency and narrative competency. ‘There is no doubt that advances associated with language and concept acquisition will transform perceptual experience, and specifically along lines that are pragmatic and intersubjective, some of which are already traced out in early non-conceptual experience’ (p.538).

But it is not clear *how* this works. This is probably why Gallagher stresses that even creatures without much experience, infants being the paradigm example, already display the kinds of skills that are representative for smart perception. According to Gallagher, infants have an *inborn drive* for social interaction.

However, we should be very careful with such an appeal to innateness. The fact that something might be innate does certainly not imply that we do not have to come up with a proper *explanation* of it. Literally, the demarcation line for what is innate and what is not depends on the instance of birth. However, although birth marks a fundamental transition point in the infant’s development, the exact timing is relatively arbitrary - the proper time of

[82] Gallagher account of direct perception has many similarities with Gibson’s (1979) account of direct perception. But Gallagher also remarks that, despite the fact that he favors a Gibsonian-style account of perception without inference or representation, he does not deny that the organism has something to contribute to the shaping of perceptual information. In other words, Gallagher’s strategy is to show how we can follow Gibson’s lead and deny the necessity of inference, while at the same time allowing for internal processing to explain how we perceive environmental properties.

birth has a bandwidth of some weeks. Besides, the infant's development does neither start nor stop there. 'Innate' does not stand in opposition to 'development'; it only indicates an earlier development in the womb. And even in this stage, the development of the fetus is not a stubborn mechanical unfolding: it depends on 'favorable circumstances' and the fetus can be severely handicapped if it is deprived of these. Even a fetus cannot be regarded separately from the special 'environment' it interacts with. In other words, to invoke the magical word 'innateness' does not free one from the job of explaining the developmental 'how'. Neither can it be taken to insure some monadic individual capacity, for the development prior to birth is just as much a relational process as it is after birth. Therefore, I think that De Jaegher (2009) is correct in her observation that 'working out a detailed account of social interaction's role in interpersonal understanding is the central element of the story of social cognition. It will allow the issue to move away from the terms of the debate set by TT and ST and followed by direct perception [...] and towards a story that explicitly connects meaning and social interaction' (p.538).

A clear focus on development helps us to take into account the phenomenological fact that social understanding in fact can be *difficult*. The interpersonal abyss as assumed by TT and ST certainly does not do justice to our pervasive experiences of mutual contact and immediate understanding of the persons close to us. On the other hand, social understanding is not always smooth and direct. This may be easier to appreciate if we take into account that, from a developmental perspective, social misunderstandings are not considered to be essentially problematic. Rather, they offer crucial *opportunities for learning*. Social learning takes place in especially those situations in which our perception is not direct, and where we are uncertain of how to proceed. De Jaegher (2009) suggests this when she says that 'Failures in understanding another's behavior are not exceptional. On the contrary, they form part and parcel of the ongoing process of social understanding. More even, misunderstandings are the pivots around which the really interesting stuff of social understanding revolves. In these instances where coordination is lost, we have the potential to gain a lot of understanding' (p.540). Discontinuity in social interaction thus leads to learning and eventually opens up new venues for intersubjective understanding.

5.

Linguistic Development and Narrative Practice

It is through hearing stories about wicked stepmothers, lost children, good but misguided kings, wolves that suckle twin boys, youngest sons who receive no inheritance but must make their own way in the world and eldest sons who waste their inheritance on riotous living and go into exile to live with the swine, that children learn or miss-learn both what a child and what a parent is, what the cast of characters may be in the drama into which they have been born and what the ways of the world are. Deprive children of stories and you leave them unscripted, anxious stutterers in their actions as in their words.

- Macintyre 1981

The linguistic turn

Embodied and embedded practices provide children with a shared context in which they learn to interpret others in terms their intentions, actions and gestures - thus enabling a basic form of social understanding. However, in our everyday life we frequently have to deal with more complex social situations in which we need more than our basic perceptions, emotions and embodied interactions. How do we get the more subtle and nuanced understanding of why people do what they do? Do we require a folk psychological theory in order to understand what they mean? Or do we need to put ourselves in their shoes and run a simulation? Although we already have explained how basic embodied and embedded practices facilitate our default and pervasive modes of social interaction, so far we haven't paid explicit attention to 'the elephant in the room': language. The development of language does not only depend on the practices described in the previous chapters, but it also carries them forward and puts them into service in much more sophisticated social contexts. As Bavidge and Ground (2009) aptly put it, 'Language changes everything [...] Language does not just make a linear difference, as it were, forward or upwards. Rather its effects wash back over activities and capacities that are not themselves in origin or nature intrinsically linguistic' (pp.26-7).

The central aim of the current chapter is to describe this 'linguistic turn' from a pragmatic perspective, and explain how it contributes to our encounters with other minds. Instead of claiming that intersubjectivity should be modeled on *individual perception*, as is done by many proponents of TT and ST, I propose that knowledge of self and others emerges with the development of actual *linguistic* competence and performance (section 1). This implies that the distinction between mind and world is an 'ontogenetic achievement', to use a phrase from Cussins (1990). Early linguistic abilities are essentially grounded in second-person interactions, and they allow us to employ a language that is publicly shared with other fellow human beings. In particular, they enable children to participate in *narrative practices*, through which they learn to put persons and contexts together in ways that allow for a much more fine-grained understanding of themselves and others (section 2). One important function of narrative is that it makes it possible for children to articulate and explicate the phenomenological content of their experiences and those of others. Narratives (unlike theories) are about individual agents, and they convey the 'what it is like' for someone to have a particular experience. But narratives have another function as well: they pull children up into the logical space of reasons, and teach them what it means to act for a reason (section 3). Initially, children's capacity to interpret others' actions in terms of reasons is severely restricted, in the sense that it is only applied successfully in rather straightforward *factive* contexts. But the acquisition of the concepts of belief and desire eventually enables them to vastly expand and improve their interpretation abilities by opening up new ways of *individuating* or *particularizing* the reasons of other agents, in a way that is tailored to the latter's psychological make-up (section 4).

5.1 Thinking in our natural language

To get an initial understanding of how language contributes to our intersubjective encounters, it helps to contrast the pragmatic view I propose in this chapter with the Cartesian view endorsed by TT and ST. According to the latter, intersubjectivity is primarily the personal achievement of an individual agent who has acquired the ability to mindread, i.e. to take a third-person theoretical stance towards others in order to predict and explain their behavior. ST argues that this ability crucially involves an *analogical* argument (and sometimes introspection as well). TT rejects the element of analogy, and claims that our understanding of others is primarily a matter of theoretical *inference*. Since both abilities presuppose mental concept mastery, however, a more basic question is how these concepts are acquired and where they get their meaning. According to theory theorists, the contents of mental concepts are fixed by their role in a theoretical network. Simulation theorists, on the other hand, claim that the contents of mental concepts are first and foremost ‘given’ to us in our own experience. But when it comes to the question of acquisition, both TT and ST seem to presuppose that mental concepts and contents are carried along by some kind of innately acquired, private language. This special language is not seen not as a product, but as a *precondition* for successful social interaction. Before we start interacting with others by means of an ‘outer’ language, we are already in possession of this ‘inner’ language.[83]

My pragmatic account, by contrast, favors what has been called the ‘Thinking in Natural Language Hypothesis’ (Davies 1998). It proposes that mental concepts and contents have to be modeled on the *actual linguistic practices* that are characteristic for more advanced forms of second-person interactions. This basically means that I tread in the footsteps of Sellars, which might seem odd since he is frequently presented as the grandfather of TT. It is indeed true that the myth of Jones seduced theory theorists into thinking that our first-person vocabulary has to be modeled on a third-person folk psychological theory. At the same time, however, there is an important difference between Sellars and his TT disciples. Unlike most proponents of TT, Sellars was quite sensitive to the importance of second-person interactions. He stressed that one of the aims of the myth of Jones was to help us to understand that ‘concepts pertaining to [...] inner episodes are primary 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’ (1956, p.107).

[83] This argument is primarily directed at *internalist* versions of TT and ST. Their externalist counterparts may be able to avoid it, but they usually lack a developmental story *altogether*. Then there are also positions that reject the requirement of mental concept mastery, but it is questionable whether they are able to articulate a suitable notion of theory and/or simulation.

This passage teaches us something important. Many theory theorists follow Sellars in modeling the mental concepts and contents of our private vocabulary on a third-person theory. However, they remain trapped in the Cartesian paradigm insofar they still try to model this theory on the perceptual abilities of *individual* minds (with the exception of those who defend an externalist version of TT). Sellars, however, claimed that the concepts we use to describe our inner episodes, just like our theoretical concepts, are primarily *intersubjective*. They are not acquired through individual perception, but emerge through second-person discursive practice. Sellars argued that 'language is essentially an *intersubjective* achievement, and is learned in intersubjective contexts' (ibid.). And in his later works, he remarked that he wrote the story of Jones as part of his search for a 'functional theory of concepts which would make their role in reasoning, rather than supposed origin in experience, their primary feature' (1975, p.285).[84]

Mental concepts and contents are not acquired in private. Instead, they are the result of a long process of linguistically mediated *interaction*, and they depend on a public space that is shared with other human beings. Accordingly, inferential reasoning and introspection only come into being when we have learned to use the unique resources of our natural language in appropriate ways. This is not a given, but a developmental *achievement*. At the same time, it is out of the question that we are able to achieve these abilities just by ourselves. On the contrary, we are taught by *others* how to employ certain linguistic constructions, or how to introspect the inner stirrings of our own mind. These others also instruct us how to represent and reason about complex states of affairs in the world. Our natural language is able to facilitate this because it has a compositional semantics. 'Words serve as anchors that allow us to speak and defer to people in our linguistic community. As long as we have this much, we can generate a representation for an unfamiliar category using purely compositional means. That is all the compositionality we need' (Prinz and Clark 2004, p.61). The components of natural-language sentences provide us with the necessary structures needed for the more open-ended, context-invariant and systematic modes of thought.

To appreciate that the language of thought is just our natural language is to emphasize the importance of linguistic development for intersubjectivity. But to argue for the importance of this linguistic turn in *developmental* terms is already to presuppose a

[84] Sellars wrote that he tried to articulate 'the logical dependence of the framework of private sense contents on the public, inter-subjective, logical space of persons and physical things'.

linguistic turn in *philosophical* terms. For it commits one to the idea that our knowledge of mind and world cannot be construed independently from our current linguistic practices. Sellars may have been one of the first philosophers to insist that we see 'mind' as a sort of hypostatization of language. He argued that the intentionality of beliefs is a reflection of the intentionality of belief sentences, rather than conversely. Such a reversal makes it possible to understand mind as gradually entering the universe by and through the gradual development of language, rather than seeing language as the outward manifestation of something inward and mysterious which humans have and animals lack. But we should add here that this not only applies to our knowledge of the *mind*. It is also true for our knowledge of the *world*. To say that both 'mind' and 'world' can be seen as hypostatizations of language amounts to saying that both our private first-person and our theoretical third-person vocabularies emerge as the result of *second-person* linguistic exchanges. As Sellars sees it, if you can explain how the social practices we call 'using language' came into existence, you have already explained all that needs to be explained about the relation between mind and world. 'Grasp of a concept is mastery of the use of a word,' Sellars says, and in this he follows Wittgenstein who already claimed that 'meaning just is use'. According to Wittgenstein, words are not defined by reference to the objects or things which they designate in the external world nor by the thoughts, ideas, or mental representations that one might associate with them, but rather by how they are used in effective, ordinary communication. 'We are inclined to forget that it is the particular use of a word only which gives the word its meaning [...] The use of the word *in practice* is its meaning' (Wittgenstein 1953, §69). This implies that, in order to understand the meaning of a word, one has to be able to engage in the linguistic practices in which it is used.[85]

Promoting this line of thinking does not imply that mind and world are 'mere linguistic constructs' that do not exist without language. But it *does* exclude the possibility that there is a view from *nowhere*. We, as human beings, cannot articulate the notions of mind and world without in some way having to rely on the linguistic practices that make such articulation possible in the first place. As Putnam (1990) puts it, elements of what we call

[85] Consider the following classical example given by Wittgenstein (1953): 'I send someone shopping. I give him a slip marked 'five red apples'. He takes the slip to the shopkeeper, who opens the drawer marked 'apples', then he looks up the word 'red' in a table and finds a color sample opposite it; then he says the series of cardinal numbers [...] up to the word 'five' and for each number he takes an apple of the same color as the sample out of the drawer. It is in this and similar ways that one operates with words.' According to Wittgenstein, we shouldn't ask what the word 'five' means, since 'No such thing was in question here, only how the word 'five' is used' (§2).

'language' or 'mind' penetrate so deeply into reality that the very project of representing ourselves as being 'mappers' of something 'language-independent' is fatally compromised from the start. From a pragmatic point of view, the primary function of language is not that of naming a thing with an intrinsic nature of its own. Instead, language is seen as a way of abbreviating the kinds of complicated interactions between mind(s) and world which are unique to us humans. These interactions are marked by verbal utterances and the use of complex linguistic constructions. They help us to coordinate our shared activities, and provide us with the tools for coping and collaborating with other minds and worldly objects rather than representing them.

5.2 Narratives about selves and others

Defining narrative

Such a pragmatic view is actually in line with TT insofar it argues that the meaning of mental states depends on how they are used in a larger conceptual framework. But instead of interpreting this framework in theoretical terms, it claims that the context-sensitive, nuanced and sophisticated nature of this framework is better captured by the notion of *narrative*.

An important feature of narrative is its concern with the *concrete* and the *particular*. This is where it importantly differs from a theory. According to proponents of TT, as we saw, our understanding of others is facilitated by a folk psychological theory that deals with the universal - it abstracts away from particular contexts towards descriptions of the way the world tends to be *in general*. If Bruner (1986) is right, a narrative does exactly the opposite: it takes context to be primary in the determination of meaning, since it deals with *specific* situations. A narrative is always *situated:* it has to be interpreted in light of a specific discourse, in order to cue interpreters to draw inferences about a structured time-course of particularized events. According to Herman (2007), 'narrative traces paths taken by particularized individuals faced with decision points at one or more temporal junctures in a story world; those paths lead to consequences that take shape against a larger backdrop in which other possible paths might have been pursued, but were not' (p.10). As a result, a narrative framework has the potential to offer a kind of practical or applied understanding of behavior that functions very differently from a theoretical one.

Another important aspect of narrative is its *temporal* structure. The *internal* time frame of a narrative reflects the serial order in which the particular events follow each other. However, for a narrative to obtain there must be more than just a temporal sequence into which events are slotted in a particular way. The events must also be such that they introduce *disruption* or *disequilibrium* into the narrated world. To be categorized as a narrative, an event-sequence must involve some kind of noteworthy disruption of an initial state of equilibrium by an unanticipated and often untoward event or chain of events. At issue here is what Bruner (1991) characterized as the dialectic of 'canonicity and breach': 'to be worth telling, a tale must be about how an implicit canonical script has been breached, violated, or deviated from in a manner to do violence to [...] the "legitimacy" of the canonical script' (p.11). Herman (2007) suggests that such a disruptive event can be seen as the *motor* of narrative, and argues that narratives prototypically follow a trajectory leading from an initial state of equilibrium, through a phase of disequilibrium, to an endpoint at which equilibrium is restored (on a different footing) because of intermediary events. Narratives display a competition between 'discordance' and 'concordance', to use Ricoeur's terminology (cf. Ricoeur 1984, 1992). On the one hand, each event in a narrative is new and different. But on the other hand, each event is part of a more general series – determined by what came before and constraining what is yet to come. It is precisely this configuration that allows the story to advance, and makes possible the basic structure of a narrative: the plot. Therefore, if we are to understand a narrative, we have to be able to identify the specific events that make up this structure, and consider connections between these events that are more than just of a temporal nature. As Roth (1991) suggests: 'Narratives give [events] a connection which is not merely chronological. The process of presenting a narrative about one's past [or the historical past] requires identifying which events are important and why' (p.178).

Besides an internal time frame, narratives are also characterized by an *external* temporality that defines the relation between the events of the narrative and the narrator who presents them. This relation might be left unspecified, something which happens in the classical type of narrative that open with the famous words 'Once upon a time...' But even in these cases, the temporal relation is usually open to a specification that these events happened in the past, or that they have not yet happened but will happen in the future, relative to the narrator's present. This is necessarily true for *self-narrative*, in which events that never happened and never will happen (fictional events) still have a specifiable place in time relative to the narrator. Gallagher (2003b) argues that ‘even if the event in question never did happen (for example, an event falsely remembered) or never will happen (for example, a planned event that never comes to fruition), in self-narrative it is still set in a temporal relation to the narrator’ (p.414).

The external time frame of a narrative is defined relative to the narrator who exists in the present. This is what provides the narrative with perspective and gives it a recognizable ‘face’. It also explains why a narrative can be characterized in terms of its ‘foregrounding of human experientiality’ (Fludernik 1996). Narratives are about particular agents and affairs that are typically human – they convey the experience of living, and are prototypically rooted in the lived, felt experience of human beings who are interacting in an ongoing way with their cohorts and surrounding environment.

Entering narrative practice: requirements and achievements

Although narrative is a practice that is specific to humans as a species, there is no need to postulate that children are *innately* disposed to tell stories. The ability to use narrative as a means for social understanding is very much dependent on and shaped by the second-person practices described in the previous chapter. However, there are also additional developmental stepping stones that must be in place in order for children to participate in narrative practice.

In the first place, children need to master the narrative’s internal time frame that reflects the serial order in which the particular events follow each other. This ability emerges by the first year, when children gradually begin to distinguish between past and future. They start to remember dynamic events, so-called *scripts*, and begin to understand sequences of familiar repeated events that involve several related actions (Bauer 1996; Bauer et al. 1994, 2000). A study by Bauer and Mandler (1990), for example, showed that 1-year-old children are already able to remember brief sequences of novel events (2 or 3 actions) over several days. And this rapidly improves when they get older; by the age of three, children can verbalize a larger number of familiar scripts in a reliable sequence (cf. Nelson and Gruendel 1981; Friedman 1991, 1992). But scripts do not yet qualify as *narratives*. They are mainly based on the child's experience of the here-and-now, and still very much lack in temporal dimension. Until their second year, the only temporal differentiation that children are capable of making is that between the present activity and everything else that has been experienced and memorized: sequences of events, people and their routines, or of places and associated objects.[86]

In order to generate a narrative, children not only have to recollect the specific past time when an event occurred, but they also have to be able to *attribute* this event to themselves or others. According to Gallagher (2003b), the first-person pronoun 'I' serves as the most minimal referent around which experienced events can be organized, and the precise way in children learn to use it (starting at around 12 months) gives them an 'extremely secure anchor' for the construction of a self-narrative. The first-person pronoun is not just a 'deflated pronoun, grammatical structure or piece of vocabulary', however. On the contrary, it has an 'embodied referent'. Gallagher argues that its use depends ontogenetically on the minimal self (cf. chapter 4.3).[87]

Both the capacity for temporal integration and the ability to self-refer by means of the first-person pronoun are necessary for the proper functioning of autobiographical memory, which provides the prior knowledge out of which a coherent self-narrative is formed.[88] It has been claimed that 2-year-old children already posses autobiographical memory. Howe (2000), for example, argues that despite the fact that the autobiographical memories of children around this age have to be elicited by questions and prompts, 'by 18-24 months of

[86] This indicates that children have not yet fully mastered the internal time perspective, which depends on the temporal integration of the sensory information in behavioral and linguistic sequences (intermodal binding). And this in turn requires a further development of working memory (WM). Neuroscience suggests that in particular the prefrontal cortex is involved in WM processes. For example, it has been shown that prefrontal cortex activity is both modulated by active memory load (Braver et al. 1997), and sustained throughout the period over which information must be maintained (Cohen et al. 1997, Courtney et al. 1997). In young children, however, the prefrontal cortex is not yet fully developed.

[87] According to Gallagher, using the first person pronoun also provides one with what Shoemaker (1984) called 'immunity to error through misidentification relative to the first-person pronoun'. When I use the first-person pronoun 'I' to refer to myself, I cannot be mistaken about the person to whom I am referring. It would be nonsensical to ask: 'Are you sure it is *you* who has toothache?' (cf. Wittgenstein 1958, p.67).

[88] Neuroscience suggests that almost all regions of the brain are involved in memory, and that episodic memories are distributed throughout the neocortex (cf. Fuster 1997). Moreover, neuropsychological studies of brain-damaged subjects show that the hippocampus, the medial temporal cortex and the prefrontal cortex play an important role in the construction of episodic memory (cf. Fletcher 1997).

age infants have a concept of themselves that is sufficiently viable to serve as a referent around which personally experienced events can be organized in memory […] the self at 18-24 months of age achieves whatever 'critical mass' is necessary to serve as an organizer and regulator of experience […] this achievement in self-awareness (recognition) is followed shortly by the onset of autobiographical memory' (pp.91-2).

An important indicator for this achievement in self-awareness is the so-called 'mirror test'. In this test, the infant is surreptitiously marked on a region of its face (that cannot be seen without the aid of a mirror), and subsequently exposed to a mirror. The idea is that, if the infant recognizes itself in the mirror, it will react by touching and exploring the marked region on its own face. By 24 months, the ability to demonstrate appropriate mark directed behavior is present in most infants (Amsterdam 1972, Bertenthal and Fischer 1978, Lewis and Brooks-Gunn 1979). This form of self-recognition is also associated with the possibility of embarrassment for having done something the wrong way (cf. Lewis 1997). Faced with a new person, children may now hide behind mother's back, for example, peeking out and back again. This is a different kind of reaction from the fear of strangers expressed at 7 or 8 months. It indicates a more objective awareness of self that is uncertain about how to behave in the presence of strangers.

Self-narrative and perspective taking

Gallagher (2003b) argues that the ability to construct a self-narrative has a certain *primacy* in shaping our understanding of self and others. He claims that 'although my own self-narrative is greatly influenced by what others say about me, and is more generally constrained by the kinds of things that *can* be said, and that *are* said about persons in my culture, it has, from a first-person perspective, a priority in shaping my self-identity. What someone else says about me will have an effect on my self-identity, and will *matter*, only if it is something that I can recognize as applying to me, and only to the extent that it fits, positively or negatively, into my own self-narrative' (pp.413-4).

According to Gallagher, the creation of a self-narrative is possible only if we are capable of using the first-person pronoun, which in turn depends on the basic sense of differentiation between self and non-self that is provided by the minimal self. Without such a differentiation, it is impossible for us to refer to ourselves with any specification, and this means that we do not have a starting point for self-narrative.

I already remarked in the previous chapter that Gallagher places much emphasis on the importance of the first-person perspective in his articulation of the minimal self. He claims, for example, that 'the minimal (or core) self possesses experiential reality, and is in fact identified with the first-person appearance of the experiential phenomena' (Gallagher and Zahavi 2008, p.204). Support for this claim is found in the quality of 'mineness', an experiential feature that stays constant throughout all experience and does not depend on something apart from the experience itself. Thus, we read that 'if the experience is given in a first-personal mode of presentation for me, it is experienced as my experience, otherwise not. In short, the self is conceived as the invariant dimension of first-personal giveness in the multitude of changing experiences' (ibid.).

The problem is that such an articulation of the minimal self comes dangerously close to one of the driving ideas behind ST: that self-understanding comes first, and can be used as a foundation for our understanding of others. As a consequence, it seems that we remain stuck in the first versus third-person debate. But there is another problem as well. Hutto (2008b) points out that one of the conditions for the possibility of recognizing that one has a point of view is that one is (potentially) able to recognize and contrast it with other points of view. It seems that one can only understand what it is to have and adopt a first-person perspective when one has learned to operate with concepts that are only made available in a second-personal social space. At the same time, however, it also seems right to say that one can have experiences even if one does not *know* it. A creature can experience even if it lacks the *concept* of experience. Therefore, the claim that one cannot recognize or understand what it is to have first-person experiences unless one is able to operate with the appropriate concepts does not preclude the having of non-conceptual feelings or experience *per se.*[89] There are many sorts of experiences that one might have

[89] This is because, as Sellars (1963) makes clear, there is a distinction between 'knowing what X is like' and 'knowing what sort of thing an X is.' The latter involves being able to link the concept of X up with other concepts in such a way as to be able to justify claims about X's. On Sellar's view, we cannot have one concept without having many, nor can we come 'to have a concept of something because we have noticed that sort of thing'; for 'to have the ability to notice a sort of thing is already to have the concept of that sort of thing (p.176). But how is a pre-linguistic child able to know what pain is, for example, if knowledge is mainly a linguistic affair? What, then, is it to know *what* pain is like without knowing or noticing what *sort* of thing it is? It is just to *have* pain. According to Rorty (1979), the snare to avoid here is the notion that 'there is some inner

prior to mastering the concept of experience. Nevertheless, these considerations raise questions about our justification for characterizing these in terms of feelings of ‘mineness’ or ‘first-personal givenness’.[90] This is why Hutto (2008b) asks: ‘For what entitles us to employ these sorts of characterization in describing the felt character of such experiences to experiencers who lack the ability to make the relevant conceptual distinctions?’ (p.15).

What I would suggest here is that non-conceptual feelings and experiences have the potential to be *expressed* – they can be articulated as soon as children have mastered the relevant linguistic capacities. This is precisely what happens when they start participating in narrative practices and learn to frame themselves and other persons in terms of narratives. However, Hutto is absolutely right that the articulation of a first-person perspective crucially depends on the possibility to recognize and contrast this perspective with those of *others*.

A closer look at the developmental evidence seems to confirm this. Children’s ability to explicitly self-attribute past events develops very much in tandem with their attribution of events to *others,* and the growing recognition that these others may have perspectives that are different from that of their own. As Nelson (2003) makes clear, there is only ‘a gradually emerging understanding of different perspectives on the world of experience, perspectives that are revealed especially in narrative discourse and that are not discernable in actions alone’ (p.29). In fact, 2-year-olds are still largely incapable of differentiating their narratives as to the source of their origin, and they usually fail to articulate and explicate the relation between the events in the narrative and the narrator who presents them. Their script-like stories still lack *perspective* - they are not yet individuated in the sense of being owned or differentiated from the stories of others who shared the experience. Nelson argues that, at this stage, narratives are ‘not yet personal or autobiographical because they are not differentiated from a nonspecific past and a social generalized world. They are stories based on the child’s life experience, but they are

illumination which takes place only when the child’s mind is lighted up by language, concepts, and descriptions, and propositions, and does not take place when the child inarticulately wails and writhes. The child feels the same thing, and it feels just the *same* to him before and after language learning. Before language learning, he is said to *know* the thing he feels just in case it is the sort of thing which in later life he will be able to make non-inferential reports about’ (p.183, italics in original).

[90] Hutto (2008b) remarks that experiences may have owners (and also that they may even have owners, *necessarily*), but argues that the question whether the owners of these experiences experience them as being *owned* is a totally different question.

no more personal than any other story' (p.31). For the young child, there is only one reality, one that is shared with others, but it is not (yet) distinctly its own. Nelson gives the example of Emily, a little girl (32 months) who reports an episode from her father, who cannot run in a marathon although he wants to. Emily puzzles about why that is.

'Today Daddy went, trying to get into the race but the people said no so he, he has to watch it on television. I don't know why that is, maybe 'cause there's too many people. I think that's why, why he couldn't go in it [...] So he has to watch it on television [...] on Halloween day, then he can run a race and I can watch him. I wish I could watch him. But they said no no no. Daddy Daddy Daddy! [...] No no, no no. Have to watch on television. But on Halloween Day he can run, run a race. Tomorrow (he'll) run (???). He says yes. Hooray! My mom and dad and a man says "you can run in the footrace," and I said "that's nice of you. I want to." So next week I'm going to [...] run to the footrace and, and run in the footrace 'cause they said I could' (Nelson 1996, p. 198).

Nelson suggests that, at this point in her development, Emily's life begins to expand beyond her own experiences and into a world that she does not know and cannot predict or explain. However, she still lives primarily in the here and now of her own understood routines. The example shows that, although Emily begins by telling a story about her father, she eventually adopts the story as though it was her own. In other words, Emily is not yet fully able to distinguish the different perspectives in the story she is telling. This is in line with other evidence that indicates that children of 2-4 years often appropriate someone else's story as their own (Miller et al. 1990), and accept (false or true) suggested aspects of episodes, or even whole episodes as being true of their own past (Ceci and Bruck 1993, Thompson et al. 1997, Bruck and Ceci 1999). During development, they only gradually move from the contribution of one or more bits of information about a certain experience to a more equal co-construction of a narrative account of this experience. As Fivush (1994) points out, between 2-5 years, the vast majority of the evaluative component of the narrative comes not from the child, but from the parent. This not only shows that children's self-narrative is importantly shaped and given form by others, but also suggests that the interactions with caregivers are crucial to their development of perspective taking. Although children of this age are already capable of shared attention, i.e. of imagining the *perceptual* perspective of the other, they still have to learn what it means to have a *narrative* perspective.

Active interpretation versus passive introspection

In the construction of narrative, (auto)biographical memory provides the background knowledge out of which a coherent narrative is formed. It is often assumed that this is simply a matter of 'encoding' and 'retrieving' information. However, the creation of a narrative is also a (*re)constructive* process – it does not merely depend on the proper functioning of memory but in an important sense contributes to the functioning of that memory. Gallagher (2003b) suggests that in order to form a narrative, 'one needs to do more than simply remember life events. One must see in such events a significance that goes beyond the events themselves; to reflectively consider them, deliberate on their meaning, and decide how they fit together semantically' (p.419). He argues that this interpretation process is facilitated by what he labels 'our meta-cognitive capacities', which allow us to fit (and sometimes force) our memories into a narrative structure. This process is guaranteed to generate a lot of *confabulation.* 'It is not unusual to construe certain events in a way that they did not in fact happen, for the sake of a unified or coherent meaning. Self-deception is not unusual; false memories are frequent. To some degree, and for the sake of creating a coherency to life, it is normal to confabulate and to enhance one's story' (ibid.).

Much is still unknown about the embodiment of our meta-cognitive capacities. Gazzaniga (1988, 1995) has suggested that they depend on a specific left-hemisphere mechanism, the so-called 'interpreter'. He argues that 'human brain architecture is organized in terms of functional modules capable of working both cooperatively and independently. These modules can carry out their functions in parallel and outside of the realm of conscious experience. The modules can effect internal and external behaviors, and do this at regular intervals. The interpreter considers all the outputs of the functional modules as soon as they are made and immediately constructs a hypothesis as to why particular actions occurred. In fact the interpreter need not be privy to why a particular module responded. Nonetheless, it will take the behavior at face value and fit the event into the large ongoing mental schema (belief system) that it has already constructed' (1988, p.219).

Gazzaniga points out that in certain cases of pathology the interpreter completely fails to integrate the behavior in a larger schema. This is clearly illustrated in experiments with split-brain patients. One of these patients, identified as J.W., still had sufficiently verbal ability in the right hemisphere to be able to understand and follow simple instructions. When the word laugh was flashed to the left visual field, and so to the right brain, he would often laugh. Prior study had determined, however, that his right brain was not sufficiently verbal to process and understand sentences or even make simple categorizations. Thus, when the investigators asked him why he had laughed, it was clear that any response to this sophisticated query would necessarily have to come from the left brain. What J.W. said was 'You guys come up and test us every month. What a way to make a living.' Apparently, the left brain developed an on-the-fly interpretation of the laughter by finding something funny in the situation and claiming that this was the cause of his behavior. In another example, the instruction 'walk' presented to the right brain resulted in the patient's getting up to leave the testing van. On being asked where he was going, the patient's left brain quickly improvised, 'I'm going into the house to get a Coke' (Gazzaniga 1983). When the interpreter functions normally, however, it tries to make sense of what *actually happened* to the person in question. In this respect, the self-understanding that results from interpretation is not completely fictional. Gazzaniga argues that it is derived 'from true facts of one's life as well as false facts that we believe to be true. The resulting spin that comes out as our personal narrative is, as a result, a bit fictional, like the idea we are in control of our behavior' (Gazzaniga and Gallagher 1998).[91]

What is problematic about Gazzaniga's story is the clear commitment to modular TT. This not only confronts us with a number of more general TT-troubles, but it also encourages a reductive explanation of narrative construction in purely neurobiological terms. However, although brain processes are without a doubt important for explaining how we are able to come to a narrative understanding of others, they would not occur

[91] Remark, however, that narratives are not only interpretations of what already has happened. Glas (2003) argues that 'The narrative is at the threshold of fact and fiction and provides, therefore, a large laboratory for moral thought experiment and the imaginary trying out of alternative life scenarios. The narrative is a way to express what one values and expects. It both presupposes and construes its own context and tradition. It both represents and construes the facts of one's life. By doing so, the narrative inscribes, with itself, the narrator in the course of a larger history. Telling is finding and anchoring one's place in the world' (p.349).

unless we were acting within a broader social context. This context has to be taken into account in order to do justice to the interactive nature of intersubjectivity. Gallagher (2003b) points out that narrative understanding should therefore be mapped out ‘on a larger and more intricate scale than that drawn in purely neurobiological accounts’, and this in turn suggests ‘an even more elaborate neurobiological picture of how [self and other understanding] is generated’ (p.419).

The challenge is to come up with a convincing story about the embodiment of narrative practice, while at the same time taking into account the much broader social and cultural context in which this narrative understanding is embedded. This section can be seen as a first step towards such a story.

5.3 Narrative practice and reason explanation [#]

Narrativity and folk psychology

Narratives enable a more sophisticated understanding of self and other because they allow us to express and articulate the experience of what it is like to be an embodied and embedded agent. But narratives are not simply about how things *are,* but also about how they *should be*. They shape the expectations we have of others (and others have of us) by making us familiar with a vast stock and wide range of ‘ordinary’ situations and the sorts of actions normally related to them. According to Hutto (2004), story-telling instills and inculcates values in children. Narratives impart norms, providing a platform from which we judge reasons and actions to be acceptable or otherwise. In the process of listening to stories, real or fictional, we learn what others will expect from us and, importantly, what we ought to expect from them. It is through narratives that we develop a properly common sense of what is ‘obvious’ and ‘significant’.

In most of our everyday intersubjective engagements, we can depend on well-rehearsed patterns of behavior and coordination, because people will do what is expected. As long as people do what they are *supposed* to do, according to the rules of social practice, they usually get along fine as ‘encultured behaviorists’. Ratcliffe (2005) suggests

[#] The sections 5.3 and 5.4 have been written in collaboration with Derek Strijbos, and I want to acknowledge him for many of the insights presented here.

that in ordinary situations, we share many of the same practices and 'canonical narratives', which tell us 'what one does', 'what should be done', 'what is to be done with artefacts of type X', or 'what those with social role Y are expected to do' in given situations. In this way narratives allow us to *directly* interpret the actions of others, i.e. without the invention of mindreading or folk psychology. Because most everyday social interaction takes place in normal (and normalized) environments, we don't have to explain or predict the behaviors of others and we don't need theory or simulation. That is why Bruner (1990) says that 'When things "are as they should be", the narratives of folk psychology are unnecessary' (p.40).[92]

But the narratives of folk psychology might come into play when the actions of others *deviate* from what we normally expect from them - when we encounter 'trouble'. This happens when we are not already familiar with the story of the other person, or when we are perplexed or surprised by his or her action. We appeal to folk psychology in situations where culturally based expectations are *violated*. Bruner (1990) argues that in these situations 'the function of the story is to find an intentional state that mitigates or at least makes comprehensible a deviation from a canonical cultural pattern' (pp.49-50).[93] The idea is that 'folk psychological narratives' can serve an explanatory function by contextualizing and normalizing behavior that is 'out of line', forging 'links between the exceptional and the ordinary' (p.47). Folk psychological narratives can smoothen our understanding of others in the cases where their actions somehow deviates from expectations and/or norms of shared practice, by revealing the *reasons* on which they acted. However, the capacity to understand actions in terms of reasons is quite sophisticated. How do we acquire this?

The narrative practice hypothesis

Hutto (2007a) argues that children enter the normative space of reasons and acquire their workaday skills in wielding folk psychology through a specific kind of second-person practice in which they are introduced to and actively engage with stories about reasons for

[92] Fodor (1987) once claimed that 'Commonsense psychology works so well it disappears' (p.3). But it seems more accurate to say that social practice works so well folk psychology is hardly needed.

[93] Bruner argues that, while a culture must contain a set of norms, 'it must also contain a set of interpretative procedures for rendering departures from those norms meaningful in terms of established patterns' (p.47).

actions.[94] This is what he calls the 'Narrative Practice Hypothesis' (NPH). The NPH focuses on paradigmatic practices of storytelling, such as children listening to and actively participating in (i.e. asking questions, being invited to make sense of the protagonist's actions, retelling the story, etc.) the tale of Little Red Riding Hood. 'The stories about those who act for reasons [...] are the foci of this practice. Stories of this special kind provide the crucial training set needed for understanding reasons. They do this by serving as exemplars, having precisely the right features to foster an understanding of the *forms* and *norms* of folk psychology' (2007b, p.53).

There are two ways in which the NPH departs radically from mainstream TT and ST accounts. First, it locates the primary origin/basis of folk psychology in *second-person*, instead of third-person, encounters. Exercising our folk psychological skills is not a 'spectator sport' of inferring reasons from actions and vice versa from a distance. The requisite 'training' takes place in conditions of mutual engagement, when people ask for and give each other reasons for their actions. Third-person prediction of action in terms of motivating reasons, Hutto claims, is a derivative and not highly reliable activity, since it necessarily involves *speculation*. As such it calls for additional third-personal resources (e.g. 'theory' or 'simulation'), which he terms 'Holmesian heuristics'. Although folk psychology can be exercised in different contexts, Hutto agrees with Bruner that most of our everyday social interactions take place in socially structured, normalized environments in which the need for action explanation is obviated.

Hutto's second departure from orthodoxy is that the NPH shifts the explanatory burden from the individual to the individual *within a socio-cultural context*. The acquisition of our folk psychological skills, he claims, cannot be properly explained by focusing on the individual in abstraction from its socio-cultural background. Advocates of TT and ST often argue that the core of our intersubjective engagements (our ability to practice folk psychology), is grounded in an internal set of principles, claiming that its acquisition is effectuated either through the biological triggering and maturation of innate folk psychological modules or through the child's private search for theoretical consistency in the social world it tries to understand. But Hutto (2004) argues that folk psychological narratives provide us with more than merely a 'framework for disinterested prediction and

[94] What is a folk-psychological narrative? Hutto nowhere makes an explicit attempt to provide a definition, and seems content to leave us with an unanalyzed and 'ordinary' understanding of this concept. Its explanatory features, however, seem to be derived from Woodward's (1984) approach to singular causal explanation.

explanation': folk psychology is an 'instrument of culture', and it gives us the grounds for 'evaluative expectations about what constitutes good reasons'.

Before children can actually play the game of giving and asking for reasons, they first have to meet a number of requirements. According to Hutto (2007a), this means that they need to have (i) a practical understanding of the propositional attitudes, and (ii) the capacity to represent the objects that these take - propositional contents as specified by that-clauses. But this is not yet sufficient, since 'having an understanding of belief is logically distinct from having an understanding of what it is to act for a reason' (p.51). Hutto argues that one can ascribe beliefs using a simple inference rule, which is useful for some social coordination purposes, such as predicting what someone might believe. However, this 'does not equate to ascribing [...] a reason: that would require ascribing [...] a complex state of mind, minimally consisting of a belief/desire pair with interlocking contents. Reasons are not to be confused with isolated thoughts or desires' (p.52). Children also need to know *how and when* folk psychology is exercised. That is, they need to acquire (iii) an understanding of the 'principles' governing the interaction of the attitudes, both with one another and with other key psychological players (such as perception and emotion), and (iv) the ability to apply all of the above sensitively (i.e. adjusting for relevant differences in particular cases by making allowances for a range of variables.[95] Hutto claims that 'proficiency in making isolated propositional attitude ascriptions -attributing certain goals, desires, thoughts and beliefs- is not the same as *knowing how* these combine to become reasons. This stronger condition must be satisfied if one is to be a folk psychologist. This requires mastery of the norms governing the interplay between these attitudes. What children are missing, even upon acquiring a practical grasp of the concept of belief, is not therefore another ingredient needed for baking the folk psychological cake - rather it is the instructions for mixing all the ingredients properly to make many such cakes' (p.53).

Hutto observes that TT and ST are conspicuously silent on the question of what grounds this practical aspect of folk psychology. This is so because 'most theorists do not accept that there is a need to give an account of such practical knowledge because they imagine, quite wrongly in my opinion, that "folk psychology" just is the name of a theory or

[95] It is important to notice that Hutto does not think that these abilities are acquired as a package deal. On the contrary, he thinks what is interesting about the NPH is precisely that it tries to explain (iii) and (iv). This is where it has something new to offer, since TT and ST do not provide a deep understanding of these abilities.

procedure; one which can be understood quite independently from its practical application' (p.33). Hutto argues that this is a serious mistake, and we only need to point to the grave problems of TT and ST in accounting for the context-sensitivity of our mindreading skills to confirm this worry. According to Hutto, folk psychology is first and foremost a *practical* enterprise that is rooted in *second-person interactions*. In order to become folk psychologically competent, we don't need to grasp a set of explicit generalizations about how others will act. Rather, we need to become familiar with the background norms for wielding folk psychology in practice, and we learn these by being exposed to the right kind of narratives. In these narratives, reasons for action are shown 'in situ', against appropriate backdrops and settings. For example, children learn how a person's reasons can be influenced by such things as their character, history, current circumstances and larger projects. In order to master the basic structure and the practical application of folk psychology, children need to be actively embedded and situated in the right kind of socio-cultural environment.

The BD-model of action interpretation

It goes without saying that I very much agree with Hutto's emphasis on the socio-cultural, practical and second-person nature of folk psychology. The question is to what extent Hutto departs from the idea of folk psychology as a theoretical affair. The NPH is definitively a huge improvement over TT and ST insofar it emphasizes that reason interpretation is primarily a *second-person practice*. In this way, Hutto seems to be able to avoid the problems of context-sensitivity and the questions about acquisition that threaten these latter positions. At the same time, however, Hutto remains committed to a *psychologized* view of action interpretation - just like his TT and ST adversaries. According to this view, understanding others in terms of reasons is primarily about the attribution of *belief-desire combinations*.

The belief-desire (BD) model of action interpretation has been close to common sense amongst theorists. Consider Currie and Sterelny (2000), for example, who state without argument that 'our basic grip on the social world depends on our being able to see our fellows as motivated by beliefs and desires we sometimes share and sometimes do not […] social understanding is deeply and almost exclusively mentalistic' (p.145-6). In similar fashion, Frith and Happé (1999) claim that 'in everyday life we make sense of each other's behavior by appeal to a belief-desire psychology' (p.2).

In some places, Hutto straightforward endorses this classical psychologized picture of action interpretation. We read, for example, that folk psychology minimally incorporates 'the practice of making sense of a person's actions using belief/desire propositional attitude psychology' (2007, p.3). Elsewhere, Hutto claims that in order to make sense of an action as performed for a reason 'it is not enough to imagine it as being sponsored by a singular kind of propositional attitude; one must also be able to ascribe other kinds of attitudes that act as relevant and necessary partners in motivational crime' (p.26). Knowledge of how the propositional attitudes interrelate with one another 'comprises what we might think of as the "core principles" of intentional psychology' (p.29).[96]

Hutto stresses that these 'principles' are not supposed to be theoretical in any meaningful sense: they do not have the form of a theory, nor are they acquired like one. At the same time, however, he just seems to take the folk psychological principles out of our heads in order to replace them by the 'principles' in our folk psychological narratives.[97] Now Hutto might object that our understanding of folk psychological narratives does not necessarily take the form of our communing with a pre-existing set of theoretical principles 'in our minds'. This is certainly true. But it also implies that, if Hutto wants to avoid the appeal to a tacit body of intrinsic knowledge, then the 'principles' he is after must (in a very explicit way) be operative in the folk psychological narratives *themselves*. Not surprisingly, Hutto thinks this is indeed the case. He boldly proclaims that 'the way beliefs and desires conspire to motivate actions - which, in abstracto, we might think of as the folk psychological schema - is a constant feature of these narratives' (2008, p.29).

So let us take a look at a concrete example. One of the best-known folk-psychological narratives that exhibits a folk psychological schema, according to Hutto, is 'Little Red Riding Hood'. He cites Lillard (1997), who tells the story as follows: 'Little Red Riding Hood *learns* from the woodcutter that her grandmother is sick. She *wants* to make her grandmother feel better [she's a nice caring girl], and she *thinks* that a basket full of treats will help, so she brings such a basket through the woods to her grandmother's house [beliefs and desires lead to actions]. When she arrives there, she *sees* the wolf in her

[96] See also Hutto (2007, p.3) where he agrees with Baker (1999) that 'belief-desire reasoning forms the core of common sense psychology'.

[97] Here a parallel can be drawn between Hutto's account and the so-called 'externalist' versions of TT discussed in chapter 1.

grandmother's bed, but she *falsely believes* that the wolf is her grandmother [appearances can be deceiving]. When she *realizes* it is a wolf, she is *frightened* and she runs away, because she *knows* that wolves can hurt people. The wolf, who indeed *wants* to eat her, leaps out of the bed and runs after her trying to catch her' (Hutto 2007, p.30, citing Lillard 1997, p.268). Hutto argues that tales of this sort are *legion*, and claims that their content and structure make them perfectly suited to teach children how the core propositional attitudes (in particular beliefs and desires) behave with respect to each other and their familiar partners: emotions, perceptions, etc.

I believe there are serious reasons to doubt this. If we take a closer look at the story under consideration, Little Red Riding Hood, then it becomes clear that the 'traditional' versions of this story (those that are told to children) actually do not contain any reference to beliefs and/or desires *at all.* Certainly, the one mentioned above does, but this is only because Lillard has inserted these references *herself*. Why? Because, as she argues, if we leave out 'our mentalistic interpretation, the tale is rather dry. A little girl hears from a woodcutter that her grandmother is sick. She walks to her grandmother's house, carrying a basket of treatments. A wolf who is in her grandmother's bed jumps up and runs after the girl. Incorporating an interpretation guided by our theory of mind makes the story a good deal more coherent and interesting' (p.268).

It is hard to see how the projection of beliefs and desires into the story of Little Red Riding Hood makes it less dry or more coherent or interesting. Consider the following, more traditional version of the story by Charles Perrault: 'Once upon a time there lived in a certain village a little country girl, the prettiest creature who was ever seen [...] One day her mother, having made some cakes, said to her, "Go, my dear, and see how your grandmother is doing, for I hear she has been very ill. Take her a cake, and this little pot of butter." Little Red Riding Hood set out immediately to go to her grandmother, who lived in another village. As she was going through the wood, she met with a wolf, who had a very great mind to eat her up, but he dared not, because of some woodcutters working nearby in the forest. He asked her where she was going. The poor child, who did not know that it was dangerous to stay and talk to a wolf, said to him, "I am going to see my grandmother and carry her a cake and a little pot of butter from my mother." "Does she live far off?" said the wolf. "Oh I say," answered Little Red Riding Hood; "it is beyond that mill you see there, at the first house in the village." "Well," said the wolf, "and I'll go and see her too. I'll go this way and go you that, and we shall see who will be there first."'

There is clearly no explicit mentioning of beliefs and desires in this version of the story. Yet, the story is coherent and interesting, and we are perfectly capable of understanding what is going on. This indicates that Ratcliffe (2008) is probably right when he remarks that things are much 'messier and more complicated' than Hutto suggests. According to Ratcliffe, it is possible to impose belief-desire patterns upon narratives such as Little Red Riding Hood if we really want to, but doing so is not very informative and it fails to do justice to the sophisticated psychological discriminations that people are able to make.

However, Hutto does seem to have some elbowroom here, since he claims that folk psychology is acquired by means of a particular kind of *education*. Folk psychology is a narrative competence that is exercised by people in *some* cultures (to varying degrees) and deployed in *certain* social situations. Thus, it seems reasonable to assume that there is considerable variety not only in people's level of folk psychological competence, but also in the narratives that are supposed to contain these folk psychological schemas. In other words, perhaps Little Red Riding Hood is just not a good example of the folk-psychological narratives we are looking for.[98]

Let us therefore consider another example offered by Hutto, this time a passage discussing Shakespeare's Othello: 'Iago intends to use Othello's positive qualities against him. What Iago means by "serve my turn upon him" is that he is going to make Othello believe that Desdemona has been unfaithful to him. The word "serve" has connotations of a prison sentence or punishment showing that Iago believes Othello deserves this cruel punishment. It also shows that Iago doesn't like him so much that he wants to personally inflict such punishment upon him even though he will personally put himself at risk he is willing to take this chance as he really doesn't like Othello. This quote is also showing that as Othello believes Iago then he does not believe in himself. He does not think that he is good enough for Desdemona as he feels that she will leave him for someone else easily' (Anonymous 2004, Hutto's italics).

[98] Hutto could also argue that folk-psychological narratives do not *explicitly* display the relations between beliefs, desires and other propositional attitudes. Instead, he could propose that they do so in an *implicit* manner: the folk psychological patterns we are looking for are potentially there, but they still have to be articulated. However, this would in turn prompt the question as to *how* children are able to do this – and is this not precisely what the NPH promised to explain? Moreover, it would reopen the door to the suggestion that children are able to recognize and identify the belief-desire structures implicit in folk psychological narratives because they *already* possess a tacit belief-desire psychology. And an appeal to tacit knowledge is probably the last thing Hutto wants. Folk-psychological narratives are supposed to explain how we acquire our folk-psychological abilities, not the other way around.

Hutto claims that what is striking about this passage is the prominent use of belief/desire terminology, and the fact that the roles for each of these attitudes appear to be pretty clearly marked out. But what does this tell us? Ratcliffe (2008) points out that there are two reasons why the Othello example is very problematic. In the first place, it is obvious that the ability to interpret sophisticated literary narratives such as Othello successfully depends upon exposure to such narratives and the appropriate training and enculturation. But this does not mean at all that our *everyday interpretation* of other people and their narratives depends on the same skills. A second and more serious worry is whether it is possible to endow the belief ascriptions found in this passage with a distinctive *explanatory role*. For Hutto does not merely claim that we can find words like ‘believes’ and ‘wants’ in the kind of narratives he promotes, but also that ‘the roles for each of these attitudes’ are ‘playing their usual parts.’ However, Ratcliffe argues that this is certainly not the case in the passage under consideration. Take the belief attribution ‘Othello believes that Desdemona has been unfaithful’, for example. According to Ratcliffe, the notion of belief serves here as convenient shorthand for something much more complicated: ‘Othello’s understanding of Desdemona’s behavior is progressively shaped by a growing sense of jealousy, distrust, emotional hurt and anger. He gradually assembles a coherent interpretation of her various activities that increasingly diverges from the reality of the situation. The relevant belief cannot be cleanly separated from the feeling that Desdemona has been unfaithful’ (2008, p.450). Ratcliffe also notices that the attribution of belief in this context also implies more than just the attribution of *information*: ‘The judgement that someone has been ‘unfaithful’ is a judgement to the effect that she or he has violated a norm, committed a betrayal, done something wrong, perhaps morally wrong. Judgements like this can serve to partially specify whether and how one ought to respond. Hence they can be motivational’ (ibid.).

Both the observation that folk psychological narratives usually lack an explicit BD-structure and the questions about the explanatory power of such a structure present potential trouble for the NPH. But I do not think that these problems are *decisive*, since Hutto is not necessarily committed to the BD-model of action interpretation. Sometimes, he even distances himself from it. For example, Hutto (2007a) advances the radical claim that ‘in understanding the reasons for which others act […] we often do not make any attribution of beliefs and desires’ (p.6). And in collaboration with Gallagher, he suggests that reasons are ‘best captured in narrative form. Coming to understand another’s reasons should not be understood as designating their discrete ‘mental states’ but their attitudes and responses as whole situated persons […] The narrative is not primarily about 'what is going on inside their heads'; it's about the events going on in the world around them, the world we share with them' (Gallagher and Hutto 2008, p.33). Such a characterization of action interpretation seems much more promising than a 'principled' one that remains loyal to the BD-model. In what follows, I aim to show that the focus on a shared world (instead of individual mental states) indeed provides us with a much better starting point for an account of reason explanation.

5.4 The primacy of second-person reason discourse

The BD-model of action interpretation was arguably inspired by philosophers such as Davidson (1963) and Goldman (1970), according to whom actions are caused by beliefs and desires. A natural consequence of this assumption is a conception of action understanding as being a form of *causal interpretation*. This explains the exclusive focus on third-person, theoretical contexts of action interpretation that is characteristic for so many TT and ST accounts: causal interpretation is typically a detached, 'sideways-on' exercise in sense-making which easily translates into reason speculation from a third-person stance in the realm of human action. Thus there appears to be a strong connection between action interpretation conceived as mental state attribution on the one hand, and a focus on third-person contexts of action interpretation on the other.

However, the effectiveness of taking such a theoretical stance towards the actions of others has been seriously overestimated. When we are perplexed by the actions of others, or try to find out what exactly motivated them to behave in certain ways, it is not clear how adopting a third-person stance and hypothesizing about their reason by means of theory or simulation will yield definite, accurate and reliable results. There are simply too many possibilities, too many reasons the agent may have out of which to select the reason she acted on. Yet this is precisely what TT and ST suggest that we do best: speculating about other people's reasons in terms of the mental states that supposedly caused the action under consideration (cf. chapter 1.2, 2.2).[99]

[99] This argument is directed at those TT, ST and hybrid TT/ST positions that explain action interpretation in terms of *mindreading* (the structural attribution of mental states such as beliefs and desires).

What the causal analysis of reason explanation appears to neglect is the default policy people follow when they are at a loss about the agent's reasons: they tend to *ask the agent* (or someone close to the agent). This is of crucial importance, however: when people engage in reason discourse, it is normally only *after* having been given the reason for the action that they can consider it, if at all, as the cause of the action. As Hutto (1999) argues: 'As long as the reason for acting is designated with deference to the agent, as opposed to giving an impersonal analysis, then labeling the reason as the cause does no discriminatory labour' (p.388).

These considerations give support to the idea that reason explanation is not a typical kind of causal explanation: in case of the latter, distinguishing an event as a cause is a condition *prior* to successful explanation. Making sense of each other in terms of reasons is prima facie not a spectator sport in which we find ourselves at a theoretical remove of others, inferring possible causes of their behavior. When we interpret other people's reasons we usually find ourselves engaged with them and their view on matters, standing face-to-face with them, asking them for their reasons, trying to follow their line of thought, asking for further information if we can't make sense of their answers and correcting them if they make a mistake. Of prime interest to the interpreter are the considerations in the light of which the agent acted or in terms of which she can rationalize her action and whether these considerations *make sense or not*, whether it is correct for the agent to follow this line of thought or not given her specific view on things within a wider practical context.

Sticking to the facts of reason discourse

If we follow this lead and widen our scope so as to include second-person contexts of interpretation, the BD-model looses much of its descriptive power. Consider the following short conversations:

A: 'Why did you come?'
B: 'Because she told me to.'

A: 'Why are you wearing a tie?'

B: 'Because it is important to make a good impression.'

A: 'Why are you getting up so early?'
B: 'It is Monday.'

According to the BD-model, these everyday reason explanations cannot be the end of the interpretation story (on the side of the interpreter), even when the interpreter understands and accepts the answers given. Reason explanations in terms of facts, values, authority of others, etc. are regarded as essentially *truncated* versions of the rationalizations the interpreter has to think through. In this way, the BD-model gives rise to a formalist picture of action interpretation, one according to which, irrespective of what has actually been said, every genuine act of action interpretation minimally take this same route of belief-desire integration. In order to really understand the answer given, to appreciate it as a genuine reason explanation, the interpreter needs to read between the lines and filter out, in all scenarios, just the right belief-desire pair to fill in the omitted syllogistic premises. But the second-person practices in which we ask and offer reasons for our actions simply do not show any sign of belief-desire attribution most of the time. In the examples mentioned above, A does not seem to be attributing specific belief-desire combo's to B. Actually, it seems there is no attribution of psychological states going on *at all*. Still, A clearly understands B's answer, yet the actual sayings give us no hints as to A's alleged attribution of specific beliefs and/or desires to B. Of course, one could argue that this is an extra tacit step that A needs take.[100] But what if A could make sense of B's action without it?

In order to provide a decent story about how this might work, consider another example. A and B have been climbing up a hiking trail, A stops for a moment to enjoy the view, while B continues on their path up the trail. Looking out into the beautiful valley below her, A suddenly sees B in the corner of her eye, walking back towards her. 'I'm coming' she says, and starts walking again. But B continues to descend in her direction, at a pretty

[100] This is indeed what proponents of the BD-model usually do when faced with these 'phenomenological' objections. They reply that the mindreading routine through which the required belief-desire pairs are constructed should be understood as a *sub-personal* process. The action principles guiding the routine are supposed to be deeply *unconscious*, and mindreading now involves having *tacit* beliefs and desires being fed into a practical reasoning *mechanism* (cf. Nichols and Stich 2003, Goldman 2006). See chapters 1.3, 2.1 and 2.2 for a critique of this strategy.

fast pace. Then she asks: ‘Why are you coming back?’ In a silent voice B answers: ‘There’s a bear!’, pointing in the direction of the bushes some fifty yards up the trail.

A wants to know why B is walking back towards her. Instead of speculating about B’s behavior, A employs the most efficient instrument at her disposal: in a rather straightforward manner she asks B why he is coming back. And B gives his reason: he is coming back *because* there is a bear (over there). According to classical psychologized accounts of action interpretation, this conversation involves a clear-cut amount of mindreading. In order for A to properly understand B’s reason, she tacitly needs to do something in the order of the following. Starting from B’s present action (his coming back) and his answer (‘There is a bear!’) together with his pointing behavior, A needs to figure out the specific mental states that gave rise to the action by calling forth a theory or running a simulation routine. In a situation such as this one, A has to recognize that B is coming back because (i) he *wants* to stay as far away as possible from that bear an (ii) he *believes* that turning back is the best way to do so (or something like this). The next step for A is to attribute this specific set of mental states to B. When this results in a similarity match with the ‘original’ motivating mental states that gave rise to B’s action, the mindreading procedure is successful, and A has properly made sense of B’s action.

However, there is a much better explanation available. It starts with the observation that B, in giving his reason for coming back, appeals to the *agent-neutral fact* that there is a bear (over there). In other words, B makes explicit that he responds to something that is the case *in their practical world*, rather than in his mind (just consider B’s pointing at *the bushes*). And this fact tells us not something about B in particular; it tells us something about the environment of *both* A and B (again, this is implicit in B’s pointing behavior). What the example suggests is that the basis for our understanding of others lies outside the mind of particular agents, in the context of a shared, factual world. When we try to make sense of the actions of others, one of the first and most important tasks is to figure out which of the facts in this world they are responding to. And the first step of the agent who is asked to explain his action in terms of reasons is precisely to provide this fact. The example clearly shows this. B makes it explicit to A that he is responding to the fact that there is a bear (over there). ‘There’s a bear’ is a ‘factive’ explanation of why he is coming back.

Human agents have a strong tendency to explain their actions with appeal to what is going on in both their and their interpreter’s environment. One might want to say that they explain and interpret their or other people's action as if what they themselves count as 'fact' were also fact for the other. But this already pays lip service to an individualist, psychologized interpretation: quite surreptitiously, this way of putting it, with its appeal to 'fact for me' and 'fact for you', renders 'fact' as something close to 'what is (justifiably) believed to be true by X'. But my point in talking about 'factive' interpretation is to stress that there are *two individuals*, one agent and one interpreter, the latter being concerned with the *former's* reason for action, but that the reason asked for and given is some state of affairs in *both's* environment.[101]

It pays to consider what Robert Gordon says about this. In his early writings on emotions, Gordon (1969) suggested that 'it is often what a person knows, as opposed to what he merely believes, that determines how his emotion is to be described' (p.409). Building on insights from Thalberg (1964) Gordon argued that the belief presupposition (S emotes that p, only if S believes that p) and the factive presupposition (S emotes that p if and only if p) that are both present in many emotion descriptions (e.g. S is annoyed that p) can be explained by a stronger presupposition: that of knowledge. The belief of S that p and the fact that p are held together in these emotion descriptions because of the implicit presupposition of S's knowledge that p. In later work this translates into the idea that many of our emotion and reason attributions to each other are not *mere* belief attributions but rather, as our factive explanations of the relevant emotions and actions suggest, implicit knowledge attributions.[102]

Importantly, Gordon notices that the factive form of interpretation is the *default* form, the one used when there is no reason not to use it. If factive interpretations indeed involve knowledge attributions, then it makes sense to claim that it is especially when factive explanations go wrong that we resort to explicit, 'mere' belief explanations: 'B is coming back because he believes there is a bear (over there). But this is exception rather than rule: 'Not, "I am doing this because I believe that p," but rather, "I am doing this because

[101] This notion of 'fact for both' that is in play here should not be analyzed in terms of 'what agent and interpreter both (justifiably) believe to be true.' The term 'both' turns up in the wrong place in the analysans, yielding belief attributions necessary for factive interpretation. And this is precisely what does not happen in cases such as the example above. Another way to say this is that the knowledge attribution implicit in factive interpretations should not be rendered as belief attribution plus something extra, or that the notion of knowledge at play here eludes definition in terms of any kind of true belief.

[102] There is a lot of empirical evidence supporting this claim, which is usually subsumed under the heading 'the curse of knowledge' (Nickerson 1999, 2001; Keysar and Bly 1995; Keysar et al. 2003; Birch and Bloom 2003, 2004).

p," or, "My reason for doing this is that p."[...] Knowledge, attributed by default, is the normal epistemic condition of others; mere belief is the noted exception' (1987, p.131-2). One could summarize Gordon's proposal as follows: in cases of factive interpretation, belief is *logically implied* by knowledge attributions without attribution of 'mere' belief being *psychologically required* to attribute knowledge.

Second person reason discourse reveals that issuing a factive explanation of an action such as 'There's a bear!' presupposes that this is a fact in both interlocuters' world. Factive reason explanations proceed under the assumption of a shared world, and this is what allows agents to explain to which fact in this world they are responding to. But I do not want to give the impression that interpretation in terms of such factive reasons may not have a proper application in third-person circumstances, when the interpreter is an observer, an onlooker. Imagine looking up from this book out the window and seeing people running for cover as it starts to pour rain from the sky. The reason why these people are seeking shelter is a fact out in the open for you to observe: it just started to rain heavily. But now suppose you can't make out what it is they are running away from, sitting behind your desk. You might walk up to the window in order to have a better look and find out that it is a bunch of water pistol fanatics aiming their weapons at their fellow students. Again, once you are in that position, you can easily discern their reason for action. No appeal needs to be made to mentalistic attribution in order to explain an observer's understanding of an agent's action in such scenarios. But at the same time no attempt should be made to reduce these cases to lower-level, quasi-behavioristic forms of interpretation where the notion of a reason is not in play.[103]

Thus, on the view presented here, default action interpretation in terms of reasons proceeds by calling upon facts in the world to which the agent responded in performing his action.[104] Of course, it sometimes happens that interpreters remain ignorant or unaware of the facts responded to. Asking the agent for his reason and being given a factive answer will then often suffice, as the example of the bear shows. But the 'facts' provided by the agent may also become subject of dispute. Obstacles of this kind are normally resolved

[103] That is: interpretation in terms of habits, socio-cultural norms or rules of conduct that would obviate the need for interpretation in terms of reasons proper.

[104] In his 'Doing things for reasons', Bittner (2001) argues that reasons are not facts, but rather states of affairs or events. Strictly speaking, facts do not occur anywhere at any time and are therefore rather odd things to respond to. I agree. Since I draw mainly from Gordon, however, I stick to his 'factive' vocabulary.

during the conversation itself, not prior to it. As I will show in the next section, unsatisfactory explanations of actions are exactly what *drives* reason discourse and often allows for a further development, or 'scaffolding' of action interpretation.

Going beyond the facts

What I propose is to view action interpretation as starting against the background of a *factual* world instead of an *individual* mind. This should not be misread as the view that interpreter and agent on default share their reasons, let alone as suggesting that agent and interpreter are not conceived as individual subjects in ordinary folk psychological practice. The interpreter interprets the agent as a subject who has *his* reason for *his* action. Yet the reason in light of which the agent acted is a feature in the practical world inhabited by agent *and* interpreter. In general, the agent's reason for his action may actually provide a reason for a similar action performed by the interpreter.

In the bear-scenario sketched above, for example, we expect A to start backing away from the bushes herself as well. Suppose A had taken better notice of B's facial expression as he was coming back towards her. She would have seen a mixture of fear and excitement that was clearly not directed at anything in the direction he was heading. Instead of asking 'Why are you coming back?' she might have asked: 'What's wrong?' B would have given the same answer: what's wrong is that there is a bear (over there)! Here the question 'What's wrong?' expresses a worry about what is going on out there, about something that might also concern A.

This version of the scenario shows that the facts responded to need not at all, and in fact rarely are, free of emotional or conative import. But such imports need not be *agent-specific.* The fact that it is raining has a certain effect on you, an effect that would make you behave in the same way as the people outside, had you been there among them. Being in the same situation together, the case of A and B and the bear is more straightforward: there being a bear over there is frightening and undesirable for A and B. But this should not tempt us to characterize the interpretation process as involving the attribution of the relevant emotions or desires to B and/or A (or to the people outside in the rain scenario). Talk of 'import' is meant to make this clear; there is something in A and B's practical world that has certain emotion and desirability characteristics.

Consider once more Gordon’s remark that default knowledge attributions are also implicit belief attributions. This is meant to make a logical point about the notion of belief (that the folk psychological notion of belief should be analyzed in terms of the folk psychological notion of knowledge and not vice versa), not a psychological point about the attribution process. Explicit attribution of belief (‘mere’ belief) normally comes with a suspension of the attribution of knowledge. I suggest analyzing imports similarly: explicit mentioning and attribution of the relevant emotions and pro attitudes is normally carried out in cases where these imports are being considered agent-specific, and thus not ‘factual’. Suppose A is an experienced bear observer. She is particularly fond of these animals and confident that she can observe them without disturbing them. Then she might as well act upon the fact that there is a bear in the bushes by approaching the bushes, leaving B behind. On passing A, B might hiss at her: ‘What are you doing?!’ A: ‘I want to have a closer look! The attribution of a desire (self-attribution in this case) makes sense precisely because the relevant fact (there being a bear over there) is responded to differently by A and B. This calls for making explicit otherwise implicit and shared imports in the guise of agent-specific desires (and emotions, cf. ‘I’m not afraid of bears’). Thus, rather than providing a basis for the possibility of a shared practice, these attributions can be conceived as emerging from A and B’s interactions in a common world.

For a similar treatment of belief attributions, consider the following scenario. It is Sunday morning, 7 ‘o clock, and C is woken up by the sound of the shower running. She tries to ignore it, but without success. After a few minutes she hears D walking back to his bedroom, next to hers. Being fully awake by now, C jumps out of bed and walks to D’s bedroom to find him getting dressed. ‘Why are you up so early’? C asks him. With surprise and being a little bit annoyed by her question, D answers: ‘Well, it’s Monday, isn’t it!?’ C: ‘No, it’s Sunday!’ D: ‘What?’ C: ‘It’s Sunday, look on your Iphone!’ D looks on his Iphone, finds out that C is right and falls on his bed, smiting his forehead. C walks back to her bedroom laughing in herself, murmuring: ‘He thought it was Monday!’

C wants to know why on earth D is getting dressed. He gives his reason by appealing to the presumed fact that it is Monday. But C knows that it is Sunday and corrects him. D realizes that there really was no reason to get up this morning and hates himself for it. Playing the sequence of events through her mind as she walks back to her room, C can’t help but laugh at the whole thing, epitomizing the joke by interpreting D’s action once again: he believed it was Monday! D acted upon something that was contrary-to-fact and therefore correctly expressed by attributing to him the *belief* that it was Monday. Attributing this belief to D is the result of an ongoing interpretation process exemplified by the piece of reason discourse cited. There is no way C could have come up with an answer to her question with such determinacy and reliability by means of speculation prior to their conversation. For D never gets his days wrong. At most, he believing that it was Monday would have been one hypothesis amongst many that could explain his behavior, a rather dissatisfying result for C (cf. Hutto 2004 for a similar point). Note that although the attribution of *belief* has a clear function in this case of action interpretation, there is no reason to think that we need to attribute a *desire* as well. Since factive action interpretations do not require us to attribute additional desires, why should belief attributions have to meet this constraint?

Now image a similar scenario, though a little bit more complicated. It's actually *Saturday*, but D thinks it is Sunday. C asks him the same question: 'Why are you up so early?' D responds that it is Sunday morning, C corrects him by making him look on his Iphone, upon which D seeks the support of his bed again. But now C asks him: 'Why would you get up at 7 'o clock if you think it is Sunday?!' D responds: 'Well, I like to go feed the birds on Sunday mornings.' C leaves his room, laughing. 'Better luck tomorrow!', she says. Apparently D provided C with enough ingredients for her to make sense of his action: he thought it was Sunday and he likes to feed the birds on Sunday mornings. Here we have both a belief and (something close to) a desire explaining D's action, just as proponents of belief-desire psychology would have it. But again it should be evident that these attributions are the result of a piece of reason discourse, and could not possibly have been made by C at the expense of this interaction. Each locution provided by the interlocuters provides a piece of the puzzle, a 'scaffolding' upon which further remarks are made and further conclusions are drawn. D expresses his belief that it is Sunday, C gives this back to him as she has another go at trying to make sense of his action. With these pieces of the puzzle in place, D answers by self-ascribing a desire towards feeding the birds on Sunday mornings, thereby providing C with enough information to follow through what moved D to act on this occasion. Notice that D's expression and self-ascription of his desire makes perfect sense here. He is well aware of the fact that feeding the birds on Sunday mornings is not something that many people are prone to. Even if it were Sunday instead of Saturday, the fact that the birds are hungry today (or something like this) is not a feature of the world that generally elicits feeding responses. But it does in the case of D. It makes sense to express and self-ascribe the desire in this case (instead of saying: ‘Because the birds are hungry today’) because we have a fact (remember it is Sunday now) that most people do not tend to respond to in D’s way: hungry birds on Sundays have an import for D that C and most other people do not experience.

Ascribing mental states in folk psychological interpretation is an act of ‘individuating’ or ‘particularizing’ of the reasons acted upon. This act of individuation reveals a reason that is particular or peculiar to the agent, in the sense that it is a reason that oneself would normally not respond to (for the interpreter, it lacks the particular affordance that it has for the agent) or that it concerns something that is or might not be the case (in the case of a false belief). But this is the exception that confirms the rule: default interpretation appeals to facts with shared imports. Return to the other example for a moment: you look out the window and see everybody running for cover as the skies unleash their fury. Your colleague enters your room with his lunch box in his hands, picking you up for your daily walk across the campus. As you look up he halts and says: ‘Ah, it’s raining!’, whereupon you reply: ‘Let’s go to the cafeteria.’ There is no need for you to attribute to your colleague the desire to stay dry, or some similar psychological attitude. The rain has the same import for both of you; the fact that it is pouring down rain is reason enough for both of you not to go outside. It is within the context of this fact to which both of you are responding that your suggestion to go to the cafeteria makes perfect sense. No process of mindreading needs to be invoked in order to explain your colleague’s understanding and subsequent endorsement of your proposal: going outside for your lunch break is out of the question, so the cafeteria starts to emerge as an appealing place to eat your lunch. This is such a common pattern in our everyday evaluation of the world around us that attribution of mental states is really beside the point. Rather, both you and your colleague can suffice by looking out into both your world and reason about the next best option for your lunch break.

Non-factive, individuated action interpretation is a more advanced and derivative form of action interpretation that comes into play when default, factive interpretation breaks down or is taken to the next level. In our everyday social practice, this moving up a level during the interpretation process is often the result of a cooperative ‘scaffolding’ of the appropriate context of action.[105] Interpretation in terms of non-factive reasons is supported

[105] This is how I would like to give shape to Hutto’s (2004, 2008) point that interpretation in terms of propositional attitudes in ordinary practice often requires second-person discourse. Not always

by the ‘scaffolds’ set up in the preceding conversation. In the examples I have restricted such scaffolding to the ingredients provided by earlier phases of the very same conversation. But in reality, scaffolding is being constructed and reconstructed throughout our (personal) lives together, in terms of our specific beliefs and desires, but also our character traits, habits, values, political and religious views, etc., depending on the intensity of our relationship together. Each conversation, each overheard remark, each action or characterization by third parties may add to the contexture of the world of others in terms of which we interpret them. But at the same time, each of these social happenings proceeds against the background of a common world we take each other to be responsive to.

How should we explain our moving up a level during the interpretation process? Here again I think that Gordon offers an interesting starting point with his notion of *ascent routines* (cf. chapter 3.2). Ascent routines allow speakers to self-ascribe propositional attitudes by redeploying the process that generates a corresponding *lower level* utterance. Goldman has objected that the chief problem for the ascent-routine approach is that this redeployment procedure can only be described for the mental classification of belief, and not for other attitudes or sensations (Goldman 2006).[106] But in a recent article, Gordon (2007) argued that ascent routines are certainly not limited to beliefs. On the contrary, Gordon claims that ‘for *every* propositional attitude type -beliefs, desires, hopes, fears, regrets, intentions, and so forth- there is a corresponding distinct ascent routine’ (p.156). He suggests that ‘the lower level utterance about “the world” and the higher level utterance about the speaker’ (ibid.) can be utterances of the *same* sentence. A 2-year-old may utter the sentence ‘[I] want a banana’ as a means of *requesting* or *demanding* a banana, *when* it

of course. Consider experimental setups for the false belief test: here attribution of a false belief is rather easy from a third-person stance for people over four years old. In such scenario’s the agent’s past interaction with the environment provides the necessary scaffolding to correctly predict her present action.

[106] Nichols and Stich (2003) endorse a stronger version of the objection: ‘We can see no way of transforming these [higher level] questions into fact questions of the sort that Gordon’s theory requires [...] There is no plausible way of recasting these questions so that they are questions about the world rather than about one’s mental state’. Another objection raised by Goldman is that Gordon’s model is in fact nothing more than a pure redeployment theory, just like Nichols and Stich’s (2003) MM account, and therefore suffer from the same debilities (cf. Goldman 2006, p.240). But Gordon replies that, unlike the MM account, the ascent routine model is not committed to the implications of a Belief (or other propositional attitude) Box, or the assumption that believing is characterized by its distinct functional role, and that these functional roles are actually implemented in the brain.

wants a banana, but without *self-ascribing* the relevant desire. If its parents would teach it to use the prefix 'I believe' whenever it makes a statement about the world, the infant would be using belief utterances when it believes something, without thereby self-ascribing the belief (which does not seem to be among the capacities of a two-year old). Gordon's improved account of ascent routines reveals that 'ascent routines can make us *reliable* self-ascribers, but they cannot make us self-ascribers' (p.164). The ascription of the attitudes requires an *embedding* of the relevant ascent routines.

This is best explained by turning back to *other*-ascriptions. In order to ascribe beliefs to others, according to Gordon, ascent routines need to be embedded in *simulations*. For example, I want to know whether someone else believes it is raining. First, I have to transform myself into the other by imaginatively occupying his situation. This involves an 'egocentric shift' or a 'recentering of the egocentric map'. Second, I ask myself, in the role of the other, the question 'Is it raining?' and my simulation links the answer to the particular individual whose situation and behavior constitute the evidence on which the simulation is based - the individual whom one is identifying with within the simulation. According to Gordon (1996), this 'gives sense to the notion of something's being a fact to a particular individual' (p.18). If the answer is affirmative, I can make the assertion 'He believes it is raining'.

Remark that there are in fact two proposals here. One is that we should understand the ascriptions of propositional attitudes (such as beliefs) to others in terms of ascent routines, which have to be exercised in proper contexts. This is mainly about the kind of *procedures* we adopt when we step up from factive to non-factive action interpretation, and move from a shared practical world towards the individuated worlds of particular agents. Another proposal is a more cognitive story about *how this might work* according to a 'radical' version of ST. On this proposal, simulation involves 'recenturing of the egocentric map'.[107]

What matters most for my purposes is the *first* proposal: belief ascriptions to others (and other attitude ascriptions) proceed by applying an ascent routine in the relevant context (the agent's particular take in the world). Deploying the concept of belief (and the other attitudes) in the service of interpretation then amounts to being able to make a semantic ascent in the context of the agent under consideration. Interpretation in terms of

[107] For Gordon, simulation is more than merely a cognitive heuristic. Rather it is something that allows us to recognize another person as someone who is 'mind-endowed' in the first place (Gordon 2004, p.2). But this is certainly not a claim I want to make.

beliefs, desires and the like is a matter of particularizing the responses of the agent beyond the facts; yet it remains a process of *looking out*, thus continuous with factive interpretation, be it now into someone's *particular* world. And referring to this process as semantic ascent makes perfect sense on my proposal: people regularly step up a level in reason discourse *when they say so*, that is, when they utter the relevant psychological vocabulary. I believe that Gordon's ascent routine model can be adopted *independently* of his *second* proposal as to *how* interpreters get at the proper context for making semantic ascent. Gordon appeals to his version of ST, but I do not think that his is the way to go in characterizing the process of scaffolded interpretation as characterized above. Neither am I confident about the appeal to 'egocentric shifts' when it comes to explaining instances of basic factive interpretation. Rather, I would like to point to the development of embodied and embodied practices described in the previous chapter.

Reason interpretation from a developmental perspective

On the BD-model of action interpretation, children are not introduced into the space of reasons before they start to get a proper hold on the propositional attitude concepts of belief and desire in the process of rationalizing the agent's action. It is widely accepted that passing the false belief test (cf. Wimmer and Perner 1983; Baron Cohen et al. 1985, Perner 1991) is a reliable indicator that infants have acquired the concept of belief. Passing this test has often been taken as the final developmental hurdle for the child's acquisition of a theory of mind. But this cannot be the end of the story according to the BD-model of action interpretation. For, as Hutto (2007a) points out, having an understanding of belief (in certain experimental setups) does not, on this model, 'equate to ascribing to X a reason: that would require ascribing to X a complex state of mind, minimally consisting of a belief/desire pair with interlocking contents' (p.26). Children must learn how these propositional attitudes and their contents interlock to form proper reasons for action. And it is unlikely that 4-year-olds have reached this level of sophistication, considering, for example, that 'research that explores whether 5-year-olds can use simple false belief knowledge to make inferences about their own and other's perspective finds that they singularly fail to do so' (Carpendale and Lewis 2004, p.91). Understanding and ascribing reasons by attributing appropriately structured belief-desire pairs apparently takes some extra years.

According to my proposal, however, there is reason to believe that children already start participating in the game of giving and asking for reason *before* the age of 4. It is quite possible that 3-year-olds are already able to understand and appreciate reasons for actions in factive contexts, since they appear to be perfectly capable of asking why-questions regarding the performances of others and understand a limited array of factive answers given in return. Admittedly, their capacity to interpret actions in terms of reasons is severely restricted in the sense that it is only applied *successfully* in rather straightforward factive contexts. But despite the fact that children of this age are not yet able to use the concepts of belief in order to distinguish between their own doxastic commitments and incompatible commitments of the agent they are interpreting, and despite the fact that they might not yet have the capacity to ascribe desires that conflict with conclusions of their own practical reasoning, they already seem to be in the position to follow through certain factive considerations of others and discern the reasons they act upon.

At the same time, children have to overcome quite a few developmental hurdles in order to exercise the capacity for factive reason interpretation. To interpret other agents as acting on certain facts that feature in a world that is fundamentally shared, children have to meet a number of important requirements. In the first place, they have to be able to respond specifically and differentially to other human beings. As I pointed out in chapter 4.1, this capacity is already operative from the moment of birth, and empirical evidence suggests that neonates and very young infants are already capable of individuating other agents and interact with them dyadically in several ways. But the infants' ability to perceive other agents as differentiated from the rest of the world clearly does not yet implicate that they also understand that there are things in this world that can be the object of *shared* attention. I argued in chapter 4.3 that such understanding only arises when infants start to participate in embedded practices and learn to interact with other agents in a triadic, world-involving way.

Yet these embodied and embedded capacities are still not sufficient for regarding the child as 'reason responsive'. The infant may be capable of discerning a limited array of means-ends relations with respect to certain rather 'proximal' goal-directed actions, and in this sense it may already have acquired sensitivity to the appropriateness or inappropriateness of certain performances by others. But this does not add up to the appreciation of the normativity of reasons proper. What I have in mind here is the normativity exposed by practical inferential patterns such as heading for the table when dinner is ready, putting your shoes on when you go outside or brushing your teeth when you go to bed. The patterns alluded to are the kinds of patterns that make up much of the stories of children's books, the patterns caregivers expose their kids to in daily life and that infants rehearse in pretend play. Such patterns reveal (possible) reasons for action *in situ*: that dinner is ready is a *reason* to head for the table, going outside is a *reason* to put on your shoes, going to bed is a *reason* to brush your teeth. Such reasonable patterns of action are also taught to the infant 'factively', in *real time*. It is *when* the infant (or someone else, e.g. the protagonist of a story) is going out, *when* it goes to bed or *when* dinner is ready that his caregiver helps him with his shoes, toothbrush or beckons him to the table. Perceptions of proximal goal-directed actions (heading for the table, putting on your shoes, brushing your teeth) get integrated into a wider view on social reality in which agents aim for goals for reasons (*because* dinner is ready, *because* you go outside, *because* you go to bed). It is especially at this factive level that Hutto's narrative practice hypothesis (2007) has real bite.

In order to appreciate the action patterns mentioned above the infant needs to type the actions and situations that are subsumed by them. Moreover, it has to learn how to articulate and attribute the (possible) factive reasons represented by these patterns. This arguably requires a fair degree of linguistic competence of the sort discussed earlier in this chapter.[108] The essentially *normative* status of these patterns captures the 'somewhat anaemic' sense in which an agent's reason for action justifies her action, namely that 'from the agent's point of view there was, when he acted, something to be said for the action' (Davidson 1963/2001, p.9). This should be taken quite literally: for the infant to get hold on reasons for action, it needs to have some understanding of *what can be said* for (or against) the agent's performances in the practice of giving and asking for reasons. It is by being introduced to and participating in this practice that the infant gets pulled up into the space of reasons, so to speak, and acquires the notion of a reason for action.

[108] Children have to be able to distinguish between past and future, and sequence actions in so-called 'scripts' in order to construct coherent and cohesive events. But they also have to learn how to use personal pronouns in order to attribute these events to themselves and to others. The development of these abilities goes hand in hand with the emergence of (auto)biographical memory and what Gallagher (2003b) calls 'meta-cognitive capacities'.

To conclude I would like to direct attention to the suggestion, implicit in the above, that interpretation in terms of factive reason is not only developmentally prior to, but also *necessary* for ascribing beliefs and desires to the agent in the course of interpreting her action in terms of reasons. My proposal can explain how such ascriptions work *in the service of* attributing reasons. TT and ST accounts often fail to acknowledge that reasons for action are given by normative creatures, in the sense specified here.[109] Consequently, they ignore the pressing question how infants are able to discern among the staggering amount of things that the agent may perceive, believe or want in a given situation, what may be *worth pursuing* in that situation. Being taught, inventing or having a nascent theory of mind that spells out how mental states relate to perceptions and behavior do not narrow down the options by far. It is not at all clear how a practically applicable notion of a reason for action could *emerge from* the infant's inventory of the psychological states of the agent. But if reasons for action are first laid out for the infant as normative patterns in a shared practical world, it is not too hard to understand what it learns at later developmental stages: to apply such normative patterns in non-factive contexts of action and make them explicit by means of ascent routines within those contexts.

To adopt this proposal is to appreciate the idea that children may already be able to make sense of others in terms of reasons before they acquire the concepts of belief and desire and apply them in the service of more sophisticated forms of reason explanation. The scaffolding of interpretation that takes place on a daily and very practical basis may then find its counterpart on the developmental level: both start from a factive baseline and proceed to more advanced forms of interpretation with the support of others.

[109] Hutto's NPH (2007), if it is indeed to be counted among them, is clearly an exception. Narratives could provide the ideal format for presenting normative patterns of action.

Epilogue

Some Consequences of Pragmatism

I have a pragmatic interest in seeing how philosophy can address issues that are not purely philosophical (at least not as purely philosophical as defined in what is, in my opinion, an overly technical and narrow sense of philosophy in the 20th century).

- Shaun Gallagher

The story so far...

The first two chapters of this book dealt with the internal problems of theory theory (TT) and simulation theory (ST) approaches to intersubjectivity. I have argued that, in the first place, both TT and ST fail to capture the interactive and relational *phenomenology* of our everyday encounters with other minds. Its proponents often parry this objection by going 'underground', arguing that the processes they postulate should be understood as being operative at the sub-personal (neurobiological or cognitive) level. In doing so, however, they implicitly seem to acknowledge that mindreading fails as an adequate characterization of intersubjectivity at the personal level. Moreover, it is questionable whether it makes sense to apply concepts at the sub-personal level that were originally coined at the personal level. Secondly, both TT and ST face serious difficulties when it comes to explaining how we are able to navigate our social environments *in the adaptive and context-sensitive way we do*. Instead, they tend to 'solve' this problem with an appeal to innateness. This, however, seems to be nothing more than an excuse for a lack of real understanding.

Chapter 3 further investigated the deeper assumptions that underlie TT and ST approaches to intersubjectivity. By accepting the problem of the other mind as a genuine problem, both TT and ST buy into a questionable picture of intersubjectivity: one that suggests a conception of the mind as a passive spectator, and takes for granted a phenomenology of uncertainty. This fosters the idea that our interactions with others require some kind of intervention between our initial observation of others and our final reaction towards them. At the center of this picture is the assumption that we are normally at a theoretical remove from other minds, and have to adopt a third-person stance towards them for the purposes of prediction, explanation and control. It is in this sense that TT and ST promote a *theoretical* approach to intersubjectivity.

By contrast, I have proposed an account of intersubjectivity that is very much *practice-oriented* (chapter 4 and 5). According to my proposal, our common sense encounters with others can be explained as being facilitated by three types of second-person practices: (i) *embodied practices*, allowing us to employ various innate or early developing capacities that provide a basic form of social understanding; (ii) *embedded practices* of joint attention, enabling an understanding of others within a broader social and pragmatic context; and (iii) *narrative practices*, providing us with stories in order to further fine-tune and sophisticate our intersubjective interactions.

These second-person practices to a large extent *obviate* the cognitively and conceptually demanding mindreading procedures postulated by TT and ST, and severely restrict the scope of intersubjective understanding in terms of mental states such as beliefs and desires. They provide us with a satisfactory explanation of our social engagements that is at the same time far more parsimonious. From a pragmatic perspective, the problem with TT and ST explanations of social interaction is that they come with severe *developmental constraints*, such as mental concept mastery, inferential abilities and analogical reasoning. If we want to take these constraints seriously (and I have argued that we should), then we cannot but conclude that young children fail to meet the necessary requirements to pursue a career in mindreading.

Another advantage of my proposal is its ability to address the TT and ST troubles with context-sensitivity. It simply points to the strong orientation towards the *concrete* and the *particular* that is characteristic for most of our interactions with other minds. At the same time, however, this presupposes a radically different notion of the mind: not as a passive, static spectator, but instead as an enactive, embodied and embedded *participant.*

All of this results in an enactive approach to intersubjectivity that increasingly works towards a *trivialization* of the problem of the other mind. It does so by challenging the four assumptions that are implicitly taken for granted by TT and ST approaches to intersubjectivity (prologue, p.15-16), arguing instead that: (i) our dealings with others are not essentially problematic; (ii) the conception of the mind that is at the basis of such a conviction should be rejected, (iii) our everyday social encounters do not *by default* require theoretical interventions, because (iv) they are firmly grounded in *second-person interactions* that can be understood in terms of direct perception-action couplings.

But does it make a difference?

Of course, the litmus test for a pragmatic second-person approach to intersubjectivity is whether it actually *makes a difference*, not only for our interpretation of the processes that facilitate social interaction, but also for empirical studies in this area. It should not come as a surprise that I think this is indeed the case. Although it is certainly not easy to investigate our social skills from a perspective that is truly second-person, doing so *will* and in fact already *has* paid off substantially. Let me give one example of a recent EEG experiment by Tognoli et al. (2007) to illustrate this claim.

Electroencephalography or EEG is a neuro-imaging technique in which a large number of electrodes are placed onto the head in order to record the electrical activity that is produced by the firings of neurons within the brain. In many EEG studies on (aspects of) intersubjectivity, lonely subjects are passively sitting in a chair while facing a monitor screen, and they are asked to perform all kinds of computer-based tasks by endlessly clicking yes or no buttons with their fingers. There is no genuine *second-person interaction* involved in these experiments. Tognoli et al. (2007), however, managed to drastically enhance and improve the set-up of their EEG experiment by placing two subjects over against each other and letting them *interact*. Initially, the subjects were asked to rhythmically wag their fingers at their own preferential pace, but they were prevented from seeing each others' hands. Then the barrier placed between them was removed, so they could see each other while continuing to wag their fingers. When subjects were allowed to see one each other's fingers moving, they sometimes adjusted their own movements and synchronized, and sometimes they did not, behaving in an independent manner. By recording, measuring and analyzing both behavior and brain activity in these interacting subjects simultaneously, the experimenters found a so-called 'phi complex', a brain rhythm operating at 10 Hz and located above the right centro-parietal cortex. According to Kelso, one of the principal investigators, these findings suggested that a unique pattern can be seen in the brains of two people interacting and that these brain activities distinguish independence from cooperation: ‘This new brain rhythm that we have discovered and termed the “phi complex” actually distinguishes when you’re socially interacting and when you’re not’.

This claim is probably highly exaggerated, and I certainly do not wish to defend my argument on the basis of the *specific* findings presented in this experiment. My only point is that this study shows that pursuing a second-person approach in scientific experimentation (and thus adopting a different conception of intersubjectivity) *does* make a difference to what we will find, even at the sub-personal level. And since such an approach does much more justice to the phenomenology of our everyday social interaction, I am convinced that it is worth pursuing.

There is another way in which my pragmatic approach makes a difference to scientific research. By stressing the irreducibility between the various practices presented in this book (the ‘levels of explanation’), it aims to discourage an interpretation of our intersubjective skills solely in terms of neurobiological mechanisms. Instead, it suggests that each level of practice might contribute to a more complete understanding of intersubjectivity. Again, an example might be helpful.

Autism spectrum disorders (ASD) are characterized by various social and communicative deficits, such as problems with imitation, empathy and language use, but also by nonsocial symptoms, such as an obsessive concern for sameness, preoccupation with objects or parts of objects, echolalia, and a variety of sensory and motor behaviors such as oversensitivity to stimuli and repetitious and odd movements (see Happé 1995, 113ff). Elucidating the underlying neural bases of ASD has been a challenge because the manifestations of this disorder vary in severity (low and high-functioning) as well as expression (Autistic Disorder, Asperger’s Disorder, and Pervasive Developmental Disorder-Not Otherwise Specified).

Nevertheless, it has been proposed that a dysfunctional mirror neuron system (MNS) early in development might be responsible for the cascade of impairments that fall under the heading of ASD (Williams et al. 2001). Despite the fact that the heterogeneity of the ASD condition seems to argue against a single cause, the idea behind this proposal is that ASD is *primarily* a failure of empathy, which in turn depends on the kind of inner imitation that is generated by the MNS (Carr et al. 2003). Now, on the one hand, studies by Dapretto et al. (2005) and Obermann (2005) suggest that there is some evidence for abnormal MNS functioning during action and observation imitation in individuals with ASD. On the other hand, however, there are also critical voices arguing that the MNS approach to intersubjectivity is seriously *flawed* (cf. Hickok 2009). Lingnau et al. (2009) have even suggested that there might be no experimental evidence for the existence of a human MNS *whatsoever.*

Of course, such a dispute by itself is not an argument against reducibility or the role of the MNS in explanations of ASD. But it does indicate that there is a serious problem with the idea that intersubjectivity = empathy = imitation = MNS, and the subsequent argument that problems with intersubjectivity therefore have to be traced back to MNS dysfunction. The problem is that the search for so-called ‘prime movers’ at the sub-personal level often implicitly results in a very *impoverished* phenomenology at the personal level, and an unjustified *simplification* of something much more complex. When such an impoverished phenomenology is used as a starting point for scientific experimentation, it yields results that are rather different compared to a much richer phenomenology that tries to capture our natural, second-person ways of dealing with other people. Klin et al. (2003), for example, has pointed out that there are remarkable differences in findings between ASD studies in which the participants were presented with static pictures of faces (e.g., Van der Geest et al. 2002), and one in which they were shown much more dynamic depictions of social interactions (by means of video). They argued that in such more ‘spontaneous’ situations, the ‘deviation from normative facescanning patterns in autism seems to be magnified’ (p.346). In other words, the attempt to replicate a more naturalistic social situation eventually gave the investigators more insight in the severity of ASD. It is highly likely that more attention for the various second-person practices in which ASD symptoms manifest themselves in the end also provides us with a *fuller* explanation of what goes wrong in ASD. Gallagher (2004) has suggested that an integrative account of ASD therefore needs to take into consideration not only possible neurobiological problems, but also dysfunctional behavior at the level of primary and secondary intersubjectivity (fig. E.1)

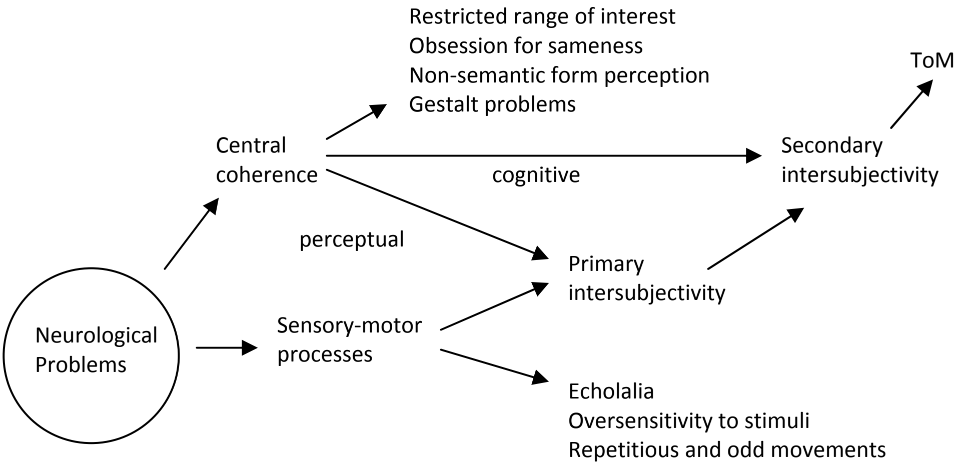

Restricted range of interest
Obsession for sameness
Non-semantic form perception
Gestalt problems
ToM
Central coherence
cognitive
Secondary intersubjectivity
perceptual
Primary intersubjectivity
Neurological Problems
Sensory-motor processes
Echolalia
Oversensitivity to stimuli
Repetitious and odd movements

Fig. E.1 A fuller picture of what can go wrong in ASD (Gallagher 2004)

Pragmatism versus reductionism

The obsession with ‘prime movers’ or ‘real causes’ that is characteristic for certain scientists clashes with a second-person approach to intersubjectivity, but it fits very well with a particular philosophical (Cartesian) paradigm. In this final section, I want to briefly comment on this paradigm and propose a view that is more in line with the pragmatic view endorsed throughout this book.

In previous chapters I have introduced a conception of the mind as a coupled complex system of brain, body and environment - one that emerges as the result of continuous interactions with other minds. I have explained these interactions in terms of embodied, embedded and narrative practices, and argued that these practices are not reducible to each other or to the sub-personal (neurobiological) processes that structure and shape them.

The idea of an emerging mind and the assumption that the practices in which it participates have their own (relative) ‘autonomy’ and explanatory pay-off is very problematic according to some philosophers. The main problem is this: if we grant the emerging mind new causal powers at each stage of development in which it grows in complexity, then how can we explain mental or ‘downward’ causation - the causal influence of a whole on its own micro-constituents?

Kim (1999), for example, argues that this kind of downward causation is either *otiose* or violates the 'causal closure of the physical' when understood to happen diachronically' (pp.28-33). This is because it relies on three principles that are mutually incompatible:

i) The physical realization principle: every emergent event or property M must be realized by (or determined by, or supervenient on) some physical event or property P (its 'emergence base').
ii) The causal inheritance principle: If M is instantiated on a given occasion by being realized by P, then the causal powers of this instance of M are identical with (or a subset of) the causal powers of P.
iii) The principle of the causal closure of the physical domain: any physical event that has a cause at time t has a physical cause at t. Hence, 'if we trace the causal ancestry of a physical event, we need never to go outside the physical domain' (Kim 1993, p.280).

In combination, these principles confront the pragmatist who is committed to emergence and believes in mental causation with a pressing dilemma: either mental causation is otiose, because the putative causal power of the emergent is preempted by the causal power of the physical elements on which the emergent is based, or mental causation violates the principle that the physical domain is causally closed.

Recently, however, Thompson (2008) has advanced a number of arguments against some of the metaphysical assumptions that underlie Kim's picture of mental causation. In what follows, I briefly discuss these arguments to the extent that they provide support for my own pragmatic proposal.

In the first place, it is important to notice that Kim accepts a 'layered model of reality', according to which the world is composed as a hierarchically stratified structure of levels of physical entities or particulars and their characteristic properties. Its bottom level consists of whatever physics is going to tell us are the most basic physical particles out of which all matter is composed (e.g. electrons, neutrons or quarks). And these objects are in turn characterized by certain fundamental physical properties and relations (e.g. mass, spin, or charm). Against this background, the challenge has become to explain how, as Kim (2000) puts it, 'it is possible for the mind to exercise its causal powers in a world that is fundamentally physical' (p.30).

Thompson criticizes this worldview because a mereologically ordered hierarchy grounded on a base level of particulars is a metaphysical picture projected onto science, whereas the image science projects is of networks of processes at various spatiotemporal scales, with no base-level particulars that ‘upwardly’ determine everything. Contemporary science does not articulate a conception of nature as grounded in a basic set of *particulars*, but instead refers to *fields* and *processes*. There is no bottom level of basic particulars with intrinsic causal properties that upwardly determine everything else. Everything is process all the way 'down' and all the way 'up', and processes are irreducibly relational - they exist only in patterns, networks, organizations, configurations, or webs (cf. Campbell and Bickhard 2002, Hattiangadi 2005).

Thompson argues that Kim’s picture of mental causation presupposes an ‘elementary-particle-version of Cartesian substance metaphysics’ that allows for part/whole reductionism. For the part/whole reductionist, 'down' and 'up' describe more and less fundamental levels of reality. Higher levels are realized by and determined by lower levels, in accordance with the layered model of reality as described in the previous section. This idea finds its expression in the principle of physical realization: every mental property M must be realized by a physical property P.

According to Thompson’s ‘process view’ of the world, however, 'up and 'down' are context-relative terms used to describe phenomena of various scales and complexity. There is no base level of elementary entities to serve as the ultimate 'emergence base' on which to ground everything. As Thompson (2007) puts it, ‘phenomena at all scales are not entities or substances but relatively stable processes, and since processes achieve stability at different levels of complexity, while still interacting with processes at other levels, all are equally real and none has absolute ontological primacy’ (p.441). Such a process view obviously fits well with the pragmatic view propounded in this book.

What about the third principle, i.e. the assumption of the causal closure of the physical domain? Thompson observes that, in the first place, it is unclear what is precisely meant by ‘physical’ in this respect. Proponents of physicalism usually talk freely of ‘mental’ and ‘physical’ properties, as if these terms track two clearly contrasting classes of entities that can be compared experimentally (cf. Strawson 2006). However, the very idea that mental properties *qua* mental can be distinguished from and systematically contrasted to physical properties in a meaningful way, as for example Mclaughlin (1994) would have it, is deeply suspect. It is simply not clear what ‘physical’ includes and excludes, and it is also hard to see how one could go about answering this question short of having a complete and true physics (cf. Montero 1999, 2001). Moreover, if we construe the principle of causal closure more narrowly as to mean the causal closure of the microphysical domain, then the principle is not obviously true and may even be false or incoherent (Dupre 1993, Hattiangadi 2005).

Although it is difficult to make intelligible the idea that complex systems are *causally* closed, Thomspon argues that there is a different way in which complex systems can said to be closed. Complex systems are closed in the sense that they are *autonomous*.[110] An autonomous system consists of a network of processes, in which (i) the processes recursively depend on each other for their generation and their realization as a network; and (ii) the processes constitute the system as a unity in whatever domain they exist.[111] According to Varela (1979), an autonomous system can be defined as a system that has *organizational closure* and *operational closure* (pp.55-60). The term 'closure' does not mean that the system is materially and energetically closed to the outside world (which of course is impossible). On the contrary, autonomous systems are thermodynamically far from equilibrium systems, which incessantly exchange matter and energy with their surroundings. 'Organizational closure' describes the self-referential (circular and recursive) network of relations that defines the system as a unity. At any given instant or moment, this self-referential network must be maintained, otherwise the system is no longer autonomous and no longer viable in whatever domain it exists. 'Operational closure' describes the recursive, re-entrant, and recurrent dynamics of the system. The system changes state on the basis of its self-organizing dynamics (in coupling with an environment), and the product of its activity is always further self-organized activity within the system (unless its operational closure is disrupted and it disintegrates).

What is important about complex systems is that they are also sufficiently open to allow for *emergent properties* with *new* causal powers. Emergent properties 'arise' out of more basic properties and yet they are 'novel' or 'irreducible' with respect to them.[112] Even

[110] 'Autonomous' literally means 'self-governing', or 'conforming to its own law'.

[111] The paradigmatic example of an emerging, self-organizing non-equilibrium system is a living cell. The constituent processes in this case are chemical; their recursive interdependence takes the form of a self-producing, metabolic network that also produces its own membrane; and this network constitutes the system as a unity in the biochemical domain. This kind of autonomy and self-production in the biochemical domain is known as *autopoiesis* (Maturana and Varela 1980).

[112] For example, Sperry (1969) writes that: 'First, conscious awareness [...] is interpreted to be a dynamic emergent property of cerebral excitation. As such conscious experience becomes

Kim (1999) does acknowledge that complex systems bring along new causal powers: 'Complex systems obviously bring new causal powers into the world, powers that cannot be identified with causal powers of the more basic simpler systems. Among them are the causal powers of microstructural, or micro-based properties of a complex system' (p.36). Strangely enough, however, he still claims that these properties are 'not themselves emergent properties; rather, they form the basal conditions from which further properties emerge (for example [...] consciousness is not itself a microstructural property of an organism, though it may emerge from one)' (ibid.) According to Kim, emergent properties such as mental properties can only be causal because they 'inherit' their causal powers from their 'emergence base' physical properties. This is what he calls the causal inheritance principle: if M is instantiated on a given occasion by being realized by P, then the causal powers of this instance of M are identical with (or a subset of) the causal powers of P (cf. Kim 1993). But Thompson argues that Kim's refusal to endow emergent properties with new causal powers is mainly due to his acceptance of part/whole reductionism, according to which micro-based properties are decomposable into the intrinsic causal properties of micro-level entities.

Of course, Thompson's story about 'mind in life' is not without problems. But it is helpful insofar it shows that some of the 'traditional' philosophical assumptions that might be in conflict with my pragmatic approach to intersubjectivity do not have to be taken for granted without questioning.[113] In this respect, the above considerations corroborate my own story about the 'mind in practice'.

inseparably tied to the material brain process with all its structural and physiological constraints. At the same time the conscious properties of brain excitation are conceived to be something distinct and special in their own right [...] Among other implications of the current view for brain research is the conclusion that a full explanation of the brain process at the conscious level will not be possible solely in terms of the biochemical and physiological data (pp. 533-5).

[113] Kim has always maintained that the problem of downward causation is primarily a *metaphysical* problem - of showing *how* mental causation is possible and not *whether* it is possible. But I think it is precisely Kim's metaphysics that is hard to swallow. Perhaps what we need is a notion of causation that is fundamentally *explanatory* (cf. Baker 1995). Instead of saying that explanation presupposes causation (as Kim does), we could say that the notion of causation presupposes a variety of explanatory practices. We do not necessarily need to motivate this skepticism about causality on Humean grounds. Norton (2003), for example, has argued that we can also justify our denial that the world is fundamentally causal by pointing at our 'enduring failures to find a contingent, universal principle of causality that holds true of our science' (p.2).

References

Amsterdam B. 1972. Mirror self-image reactions before age two. *Development Psychobiology* 5, 297-305.

Augustine A. 1991. *Confessions* (translated by Chadwick H.) New York: Oxford University Press.

Avramides A. 2001. *Other Minds*. London: Routledge.

Ayer A.J. 1956. *The Problem of Knowledge*. London: Macmillan.

Baird J.A. and Baldwin D.A. 2001. Making sense of human behavior: Action parsing and intentional inference. In Malle B.F., Moses L.J. and Baldwin D.A. (eds.) *Intentions and Intentionality: foundations of Social Cognition* (193-206). Cambridge MA: MIT Press.

Baker L.R.

1995. *Explaining Attitudes: A Practical Approach to the Mind.* Cambridge: Cambridge University Press.

1999. What is This Thing Called 'Commonsense Psychology'? *Philosophical Explorations* 2, 3-19.

Baldwin D.A., Baird J.A., Saylor M.M. and Clark M.A. 2001. Infants parse dynamic action. *Child Development* 72, 708-17.

Baron-Cohen S.

1989. The autistic child's theory of mind: A case of specific developmental delay. *Journal of Child Psychology and Psychiatry* 30, 285-98.

1995. *Mindblindness: an essay on autism and theory of mind.* Cambridge MA: MIT Press/Bradford Books.

Baron-Cohen S., Leslie A.M., and Frith U. 1985. Does the autistic child have a theory of mind? *Cognition* 21, 37-46.

Baron-Cohen S., Leslie A.M. and Frith U. 1986. Mechanical, behavioral and Intentional understanding of picture stories in autistic children. *British Journal of Developmental Psychology* 4, 113-125.

Barr R., Dowden A. and Hayne H. 1996. Developmental changes in deferred imitation by 6 to 24-month-old infants. *Infant Behavior and Development* 19, 159–70.

Bates E., Camioni L., and Volterra V. 1975. The acquisition of performatives prior to speech. *Merrill-Palmer Quarterly* 21, 205–26.

Bauer P.J. 1996. *The development of memory in childhood*. London: University College London Press.

References

Bauer P. J. and Mandler J. M. 1992. Putting the horse before the cart: The use of temporal order in recall of events by one-year-old children. *Developmental Psychology* 28, 441-52.

Bauer P.J., Hertsgaard L.A. and Dow G.A. 1994. After 8 months have passed: long-term recall of events by 1- and 2-year-old children. *Memory* 2, 353-83.

Bauer P.J., Wenner J.A., Dropik P.L. and Wewerka S.S. 2000. Parameters of remembering and forgetting in the transition from infancy to early childhood. *Monographs of the Society for Research in Child Development* 65, 1-204.

Bavidge M. and Ground I. 2009 Do animals need a theory of mind? In Leudar I. and Costall A. (eds.) *Against Theory of Mind*. Basingstoke: Palgrave Macmillan.

Baynes K. and Gazzaniga M.S. 2000. Consciousness, introspection, and the split-brain: The two minds/one body problem. In Gazzaniga M.S. (ed.) *The New Cognitive Neurosciences Second Edition* (1355-68) Cambridge MA: MIT Press.

Bermudez J.

1998. *The Paradox of Self-Consciousness*. Cambridge MA: MIT Press.

2003. *Thinking Without Words.* Oxford: Oxford University Press.

Bernier P. 2002. From simulation to theory. In Dokic, J. and Proust J. (eds.) *Simulation and Knowledge of Action* (33-48) Amsterdam: John Benjamins.

Bertenthal B.I. and Fischer K. W. 1978. Development of self-recognition in the infant. *Developmental Psychology* 14, 44-50.

Birch S. and Bloom P. 2003. Children are cursed: An asymmetric bias in mental state attribution. *Psychological Science* 14, 283-6.

Birch S. and Bloom P. 2004. Understanding children's and adult's limitations in mental state reasoning. *Trends in Cognitive Science* 8, 255-60.

Bittner R. 2001. *Doing things for reasons.* Oxford: Oxford University Press.

Blakemore S.J. and Decety J. 2001. From the perception of action to the understanding of intention. *Nature Reviews Neuroscience* 2, 561-7.

Blakemore S.J., Frith C.D. and Wolpert D.W. 2001. The cerebellum is involved in predicting the sensory consequences of action. *Neuroreport* 12, 1879–85.

Bloom P. and German T.P. 2000. Two reasons to abandon the false belief task as a test of theory of mind. *Cognition* 77, B25-B31.

Bogdan R. 1997. *Interpreting Minds: The Evolution of a Practice*. Cambridge MA: MIT Press.

References

Borg E. 2007. If Mirror Neurons are the Answer, What was the Question? *Journal of Consciousness Studies* 14, 5-19.

Botterill G. 1996. Folk psychology and theoretical status. In Carruthers P. and Smith P. (eds.) *Theories of Theories of Mind* (184-99) Cambridge: Cambridge University Press.

Braddon-Mitchell D. and Jackson F. 2007. *Philosophy of Mind and Cognition: An Introduction*. Oxford: Blackwell Publishing.

Brasil-Neto J.P., Cohen L.G., Pascual-Leone A., Jabir F.K., Wall R.T. and Hallett M. 1992. Rapid reversible modulation of human motor outputs after transient deafferentation of the forearm: a study with transcranial magnetic stimulation. *Neurology* 42, 1302–6.

Braver T.S., Cohen J.D., Nystrom L.E., Jonides J., Smith E.E. and Noll D.C. 1997. A parametric study of prefrontal cortex involvement in human working memory. *Neuroimage* 5, pp. 49–62.

Bretherton I. 1991. Intentional communication and the development of an understanding of mind. In Frye, D. and Moore, C. (eds.) *Children's theories of mind* (49-75). Hillsdale NJ: Erlbaum.

Bruck M., and Ceci S.J. 1999. The suggestibility of children's memory. *Annual Review of Psychology* 50, 419-39.

Bruner J.S.

1990. *Acts of meaning.* Cambridge MA: Harvard University Press.

1991. The Narrative Construction of Reality. *Critical Inquiry* 18, 1–21.

Bruner J. and Kalmar D.A. 1998. Narrative and metanarrative in the construction of self. In Ferrari M. and Sternberg R. J. (eds.) *Self Awareness: Its Nature and Development* (308-31) New York: Guilford Press.

Butterworth G. 1991. The ontogeny and phylogeny of joint visual attention. In Whiten A. (ed.) *Natural theories of mind* (223–32). Oxford: Blackwell.

Butterworth G. and Grover L. 1990. Joint visual attention, manual pointing, and preverbal communication in human infancy. In Jeannerod M. (ed.) *Attention and performance XIII: Motor representation and control* (605-24) Hillsdale, NJ: Erlbaum.

Butterworth G. and Jarrett, N. 1991. What minds have in common is space: spatial mechanisms serving joint visual attention in infancy. *British Journal of Developmental Psychology* 9, 55-72.

Byrne R.W. and Whiten A. 1988. *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes and Humans.* Oxford: Oxford University Press.

References

Campbell R.J. and Bickhard M.H. 2002. Physicalism, emergence, and downward causation. *Unpublished*. (URL: http://www.lehigh.edu/~mhb0/physicalemergence.pdf)

Carpendale J.L.M. and Lewis C. 2004. Constructing an understanding of the mind: the development of children’s social understanding within social interaction. *Behavorial and Brain Sciences* 27, 79–151.

Carpenter M., Akhtar N. and Tomasello M. 1998. Fourteen-through 18-month-old infants differentially imitate intentional and accidental actions. *Infant Behavior Development* 21, 315–30.

Carr L., Iacoboni M., Dubeau M.C., Mazziotta J.C. and Lenzi G.L. 2003. Neural mechanisms of empathy in humans: a relay from neural systems for imitation to limbic areas. *PNAS* 100, 5497–502.

Carruthers P. 1996. Simulation and self-knowledge: a defense of theory-theory. In Carruthers P. and Smith P.K. (eds.) *Theories of Theories of Mind* (22-38). Cambridge: Cambridge University Press.

Cash T.F. and Brown T.A. 1987. Body image in anorexia nervosa and bulimia nervosa: a review of the literature. *Behavior Modification* 11, 487–521.

Castelli F., Frith C., Happé F., and Frith U. 2002. Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain 125*, 1839-49.

Ceci S.J. and Bruck M. 1993. Suggestibility of the child witness: A historical review and *synthesis. Psychological Bulletin* 113, 403-39.

Chaminade T., Meltzoff A. and Decety J. 2002. Does the end justify the means? A PET exploration of the mechanisms involved in human imitation. *NeuroImage* 15, 318–28.

Charman T., Baron-Cohen S., Swettenham J., Baird G., Cox A. and Drew A. 2000. Testing joint attention, imitation, and play as infancy precursors to language and theory of mind. *Cognitive Development* 15, 481–98.

Chartrand T. and Bargh J. 1999. The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology* 76, 893-910.

Cheng P.W. and Holyoak K.J. 1985. Pragmatic reasoning schemas. *Cognitive Psychology* 17, 391-416.

Chomsky N. 1957. *Syntactic Structures*. The Hague/Paris: Mouto.

References

Churchland P.M.

1981. Eliminative Materialism and the Propositional Attitudes. *Journal of Philosophy* 78, 67-90.

1988. Matter and consciousness. Cambridge Massachusetts: MIT Press.

Churchland P.S. 1986. *Neurophilosophy: Toward a Unified Science of the Mind/Brain*. Cambridge: MIT Press.

Cilliers P. 2005. Knowing complex systems. In Richardson, K. (ed.) *Managing Organizational Complexity: Philosophy, Theory, and Application* (7–20) Greenwich: Information Age Publishers.

Clements W.A. and Perner J. 1994. Implicit understanding of belief. *Cognitive Development* 9, 377-95.

Cohen J.D., Perstein W.M., Braver T.S., Nystrom L.E., Noll D.C., Jonides J. and Smith E.E. 1997. Temporal dynamics of brain activation during a working memory task. *Nature* 386, 604-8.

Cosmides L. 1989. The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition* 31, 187-276.

Courtney S.M., Ungerleider L.G., Keil K. and Haxby J.V. 1997. Transient and Sustained Activity in a Distributed Neural System for Human Working Memory. *Nature* 386, 608-11.

Crane T. 2003. *The Mechanical Mind: A Philosophical Introduction to Minds, Machines, and Mental Representation 2nd Edition*. New York: Routledge.

Csibra G.

2005. Mirror neurons and action observation. Is simulation involved? ESF Interdisciplines. (http://www.interdisciplines.org/mirror/papers).

2007. Action mirroring and action understanding: an alternative account. In Haggard P., Rosetti Y. and Kawato M. (eds.) *Sensorimotor foundations of higher cognition. Attention and performance XII* (453-9) Oxford: Oxford University Press.

Csibra G. and Gergely G. 2009. Natural pedagogy. *Trends in Cognitive Sciences* 13, 148-53.

Currie G. 1995. Imagination and Simulation. In Davies, M. and Stone, T. (eds.) *Mental Simulation* (151-69) Oxford: Blackwell.

Currie G. and Ravenscroft I. 2003. *Recreative Minds.* Oxford: Oxford University Press.

Currie G. and Sterelny K. 2000. How to think about the Modularity of Mind-Reading. *Philosophical Quarterly* 50, 143–60.

Cussins A. 1990. The connectionist construction of concepts. In Boden M. (ed.) *The Philosophy of Artificial Intelligence* (368-440) Oxford: Oxford University Press.

Dapretto M., Davies M.S., Pfeifer J.H., Scott A.A., Sigman M., Bookheimer S.Y. and Iacoboni M. 2006. Understanding emotions in others: mirror neuron dysfunction in children with autism spectrum disorders. *Nature Neuroscience* 9, 28-30.

Davidson D.

1963. Actions, Reasons, and Causes. *Journal of Philosophy* 60, 685-700 (Reprinted in Davidson 2001).

1984. *Inquiries into Truth and Interpretation.* Oxford: Clarendon Press.

2001. *Essays on Actions and Events*. New York*:* Oxford University Press.

Davies M. 1998. Language, Thought, and the Language of Thought (Aunty's Own Argument Revisited). In Carruthers P. and Boucher J. (eds.) *Language and Thought* (226–47) Cambridge: Cambridge University Press.

Davies M. and Stone T. (eds.)

1995a. *Mental Simulation: Evaluations and Applications.* Oxford: Blackwell.

1995b. *Folk Psychology: The Theory of Mind Debate.* Oxford: Blackwell.

Decety J., Grezes J., Costes N., Perani D., Jeannerod M., Procyk E., Grassi F. and Fazio F. 1997. Brain activity during observation of actions: Influence of action content and subject's strategy. *Brain* 120, 1763-77.

Decety J., Chaminade T., Grezes J. and Meltzoff A. 2002. A PET exploration of the neural mechanisms involved in reciprocal imitation. *NeuroImage* 15, 265-72.

Decety J. and Jackson P.L. 2004. The functional architecture of human empathy. *Behavioral and Cognitive Neuroscience Reviews* 3, 71-100.

De Jaegher H. 2009. Social understanding through direct perception? Yes, by interacting. *Consciousness and Cognition* 18, 535-42.

De Jaegher H. and Di Paolo E. 2007. Participatory Sense-Making: An enactive approach to social cognition. *Phenomenology and the Cognitive Sciences* 6, 485-507.

Dennett D.C.

1969. *Content and consciousness.* London: Routledge and Kegan Paul.

1978. Beliefs about beliefs. *Behavioral and Brain Sciences* 1, 568-70.

1987. *The Intentional Stance.* Cambridge MA: MIT Press.

1991. Real Patterns. *Journal of Philosophy* 88, 27-51.

Descartes R.

1904. *Oeuvres de Descartes*. Adam C. and Tannery T. (eds.) Paris: Vrin.

1984. *The Philosophical Writings of Descartes Volumes I and II*. Cottingham J., Stoothoff R. and Murdoch D. (eds.) Cambridge: Cambridge University Press.

1984. *The Philosophical Writings of Descartes Volume III*. 1984. Cottingham J., Stoothoff R., Murdoch D. and Kenny A. (eds.) Cambridge: Cambridge University Press.

De Vignemont F. 2004. The co-consciousness hypothesis. *Phenomenology and the Cognitive Sciences* 3, 97-114.

Dewey J. 1960. *The Quest for certainty.* New York: Capricorn.

Dijksterhuis A. and Van Knippenberg A. 1998. The relation between perception and behavior, or how to win a game of trivial pursuit. *Journal of Personality and Social Psychology* 74, 865-77.

Di Pellegrino G., Fadiga L., Fogassi L., Gallese V. and Rizzolatti G. 1992. Understanding motor events: A neurophysiological study. *Experimental Brain Research* 91, 176–80.

Donald M.

1991. *Origins of the Modern Mind: Three Stages in the Evolution of Culture and Cognition*. Cambridge MA: Harvard University Press.

2001. *A Mind So Rare: The Evolution of Human Consciousness*. New York: W.W. Norton.

Doyle A.C. 1887-1927/1986. *The complete Illustrated Sherlock Holmes.* Chatham: Omega Books Ltd.

Dunn J. 1988. *The beginnings of social understanding*. Oxford: Blackwell.

Dupré J. 1993. *The Disorder of Things.* Cambridge MA: Harvard University Press.

Edelman G.M. 1992. *Bright Air*, *Brilliant Fire. On the Matter* of *Mind*. New York: Basic Books.

Eilan N. 2005. *Joint Attention, Communication, and Mind*. In Eilan N., Hoerl C., McCormack T. and Roessler J. (eds.) *Joint Attention: Communicationd and Other Minds* (1-33) Oxford: Clarendon Press.

Evans G. 1982. *The Varieties of Reference* (ed. J. McDowell) Oxford: Oxford University Press.

Farrer C. and Frith C.D. 2002. Experiencing oneself vs. another person as being the cause of an action: the neural correlates of the experience of agency. *Neuroimage 15,* 596-603.

Farrer C., Franck N., Georgieff N., Frith C.D., Decety J. and Jeannerod M. 2003. Modulating the experience of agency: a positron emission tomography study. *Neuroimage* 18, 324-33.

Feyerabend P. 1962. Explanation, reduction and empiricism. In Feigl H. and Maxwell G. (eds.) *Minnesota studies in the philosophy of science* Volume 3 (28-97) Minneapolis: University of Minnesota Press.

Field T.M., Woodson R. and Greenberg R. 1982. Discrimination and imitation of facial expressions by neonates. *Science* 218, 179-81.

Field T., Sandburg S., Garcia R., Vega-Lahr N., Goldstein S. and Guy L. 1985. Pregnancy problems, postpartum depression, and early mother-infant interactions*.* *Developmental Psychology* 21, 1152-6.

Fivush R. 1994. Constructing narrative, emotion, and self in parent-child conversations about the past. In Neisser U. (ed.) *The remembering self: Construction and accuracy in the self-narrative* (136-57) New York: Cambridge University Press.

Flavell J.H., Green F.L. and Flavell E.R. 1986. Development of Knowledge about the Appearance-Reality Distinction. *Monographs of the Society for Research in Child Development* 51, 1.

Fletcher P.C., Happé F., Frith U., Baker S.C., Dolan R.J., Frackowiak R.S. and Frith C.D. 1995. Other minds in the brain: a functional imaging study of 'theory of mind' in story comprehension. *Cognition* 5, 109–28.

Fletcher P.C., Frith C.D. and Rugg M.D. 1997. The Functional Neuroanatomy of Episodic Memory. *Trends in Neuroscience* 20, 213-18.

Fludernik M. 1996. *Towards a 'Natural' Narratology*. London: Routledge.

Fodor J.A.

1975. *The Language of Thought*. Cambridge MA: Harvard University Press.

1983. *The modularity of mind.* Cambridge MA: MIT Press.

1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge MA: MIT Press.

1995. A theory of the child's theory of mind. In Davies M. and Stone T. (eds.) *Mental Simulation* (33-52) Oxford: Blackwell.

Fogassi L., Ferrari P.F., Gesierich B., Rozzi S., Chersi F. and Rizzolatti G. 2005. Parietal lobe: From action organization to intention understanding. *Science* 302, 662-7.

Frankish K. 2004. *Mind and supermind*. Cambridge: Cambridge University Press.

Friedman W.J. 1991. The development of children's memory for the time of past events. *Child Development* 62, 139-55.

Friedman W.J. 1992. Children's time memory: The development of a differentiated past. *Cognitive Development* 7, 171–87.

Frith U. and Happé F. 1999. Theory of mind and self consciousness: What is it like to be autistic? *Mind and Language* 14, 1-22.

Fuchs T. and De Jaegher H. 2009. Enactive intersubjectivity: Participatory sense-making and mutual incorporation. *Phenomenology and the Cognitive Sciences* 8, 465-86.

Fuller G. 1995. Simulation and psychological concepts. In Davies M. and Stone T. (eds.) *Mental Simulation* (19-32) Oxford: Blackwell.

Fuster J.M. 1997. Network memory. *Trends in Neuroscience* 20, 451-8.

Gallagher S.

1997. Mutual Enlightenment: Recent Phenomenology in Cognitive Science. *Journal of Consciousness Studies* 4, 195-214.

2000. Philosophical conceptions of the self: implications for cognitive science. *Trends in Cognitive Science* 4, 14-21.

2001. The practice of mind: Theory, simulation or primary interaction? *Journal of Consciousness Studies* 8, 83–108.

2003a. Bodily self-awareness and object-perception. *Theoria et Historia Scientiarum: International Journal for Interdisciplinary Studies* 7, 53-68.

2003b. Self-narrative, embodied action, and social context. In Wiercinski A. (ed.) *Between Suspicion and Sympathy: Paul Ricoeur's Unstable Equilibrium* (409-423) Toronto: The Hermeneutic Press.

2004. Understanding interpersonal problems in autism: Interaction theory as an alternative to theory of mind. *Philosophy, Psychiatry, and Psychology* 11, 199-217.

2005. *How the Body Shapes the Mind*. New York: Oxford University Press.

2007. Logical and phenomenological arguments against simulation theory. In Hutto D. and Ratcliffe M. (eds.) *Folk Psychology Re-assessed* (63-78) Dordrecht: Springer Publishers.

Gallagher S. and Meltzoff A.N. 1996. The earliest sense of self and others. *Philosophical Psychology* 9, 213–36.

Gallagher S. and Brøsted Sørensen J. 2006. Experimenting with phenomenology. *Consciousness and Cognition* 15, 119-34.

Gallagher S. and Hutto D. 2008. Understanding others through primary interaction and narrative practice. In Zlatev J., Racine T., Sinha C. and Itkonen E. (eds.) *The Shared Mind: Perspectives on Intersubjectivity* (17-38) Amsterdam: John Benjamins.

Gallagher S. and Varela F. 2003. Redrawing the map and resetting the time: Phenomenology and the cognitive sciences. *Canadian Journal of Philosophy* 29, 93-132.

Gallagher S. and Zahavi D. 2008. The phenomenological mind: an introduction to philosophy of mind and cognitive sciences. London/New York: Routledge.

Gallese V. 2001. The 'shared manifold' hypothesis: from mirror neurons to empathy. *Journal of Consciousness Studies* 8, 33-50.

Gallese V., Keysers C. and Rizzolatti G. 2004. A unifying view of the basis of social cognition. *Trends in Cognitive Sciences* 8, 396-403.

Gallese V. 2005. 'Being like me': Self-other identity, mirror neurons and empathy. In Hurley S. and Chater N. (eds.) *Perspectives on Imitation* (101-18) Cambridge MA: MIT Press.

Gallese V. and Goldman A. 1998. Mirror neurons and the Simulation Theory of mindreading. *Trends in Cognitive Sciences* 2, 493–501.

Gallese V., Fadiga L., Fogassi L. and Rizzolatti G. 1996. Action recognition in the premotor cortex. *Brain* 119, 593-609.

Garfield J.L., Peterson C.C. and Perry T. 2001. Social Cognition, Language acquisition and the development of the Theory of Mind. *Mind and Language* 16, 494–541.

Gardner R.M. and Moncrieff C. 1988. Body image distortion in anorexics as a non-sensory phenomenon: a signal detection approach. *Journal of Clinical Psychology* 44, 101–7.

Garvey C. 1990. *Play.* Cambridge MA: Harvard University Press.

Gazzaniga M.S.

1988. *Mind matters*. Boston: Houghton Mifflin.

1995. Principles of human brain organization derived from split-brain studies. *Neuron* 14, 217-28.

1998. *The Mind's Past*. Berkeley: University of California Press.

1992. *Nature's Mind: The biological Roots of thinking, Emotions, Sexuality, Language, and Intelligence*. New York: Basic Books.

2000. Cerebral specialization and interhemispheric communication. Does the corpus callosum enable the human condition? *Brain* 123, 1293-326.

Gazzaniga M.S. and Gallagher S. 1998. The Neuronal Platonist. *Journal of Consciousness Studies* 5, 706-17.

Gergely G. 2001. The obscure object of desire: 'Nearly, but clearly not, like me': Contingency preference in normal children versus children with autism. *Bulletin of the Menninger Clinic* 65, 411-26.

Gergely G., Bekkering H. and Király I. 2002. Rational imitation in preverbal infants. *Nature 415,* 755.

Gergely G. and Csibra G. 2003. Teleological reasoning in infancy: The naive theory of rational action. *Trends in Cognitive Sciences* 7, 287-92.

Georgieff N. and Jeannerod M. 1998. Beyond consciousness of external events: A 'Who' system for consciousness of action and self-consciousness. *Consciousness and Cognition* 7, 465-77.

Gibson J.J. 1997. *The ecological Approach to Visual perception.* Boston: Houghton Mifflin.

Glas G. 2003. Idem, ipse, and loss of the self. *Philosophy, Psychiatry and Psychology* 10, 347-52.

Godfrey-Smith P. 2003. Folk Psychology Under Stress: Comments on Susan Hurley's Animal Action in the Space of Reasons. *Mind and Language* 18, 266-72.

Goldman A.I.

1970. *A theory of human action*. Englewood Cliffs New Jersey: Prentice-Hall.

1989. Interpretation psychologized. *Mind and Language* 4, 161–85.

1993. The psychology of folk psychology. *Behavioral and Brain Sciences* 16, 15-28.

1995. Interpretation Psychologized. In Davis M. and Stone T. (eds.) *Folk Psychology: the Theory of Mind Debate* (74-99) Oxford: Blackwell.

2000. Folk psychology and mental concepts. *Protosociology* 14, 4-25.

2002. Simulation theory and mental concepts. In Dokic J. and Proust J. (eds.) *Simulation and Knowledge of Action* (1-19) Amsterdam: John Benjamins.

2006. *Simulating minds: The philosophy, psychology and neuroscience of mindreading*. Oxford: Oxford University Press.

Goldman A.I. and Sripada C.S. 2005. Simulationist models of face-based emotion recognition. *Cognition* 94, 193–213.

Gopnik A.

1993. How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences* 16, 1–14.

2003. The theory theory as an alternative to the innateness hypothesis. In Antony L. and Hornstein N. (eds.) *Chomsky and his critics* (238-54) Oxford: Blackwell.

Gopnik A. and Astington J.W. 1988. Children's Understanding of Representational Change and its Relation to the Understanding of False Belief and the Appearance-Reality Distinction. *Child Development* 59, 26-37.

Gopnik A. and Meltzoff A.N.

1994. Minds, bodies and persons: Young children’s understanding of the self and others as reflected in imitation and "theory of mind" research. In Parker S. and Mitchell R. (eds.) *Self-awareness in animals and humans* (166-86) New York: Cambridge University Press.

1997. *Words, Thoughts, and Theories*. Cambridge MA: MIT Press.

Gopnik A. and Wellman H.M. 1992. Why the child’s theory of mind really ‘is’ a theory. *Mind and Language* 7, 145-71.

Gopnik A. and Wellman H.M. 1994. The 'Theory Theory’. In Hirschfield L. and Gelman S. (eds.) *Mapping the Mind: Domain Specificity in Culture and Cognition* (257-93) New York: Cambridge University Press.

Gordon R.M.

1969. Emotions and Knowledge. *The Journal of Philosophy* 66, 408-13.

1986. Folk psychology as Simulation. *Mind and Language* 1, 158-71.

1987. *The Structure of Emotions: Investigations in Cognitive Philosophy*. Cambridge: Cambridge University Press.

1992. The Simulation Theory: Objections and misconceptions. *Mind and Language* 7, 11–34.

1995. Simulation without introspection or inference from me to you. In M. Davies and T. Stone (eds.) *Mental Simulation* (53–67) Oxford: Blackwell.

1996. Radical Simulationism. In Carruthers P. and Smith P. (eds.) *Theories of Theories of Mind* (11-21) Cambridge: Cambridge University Press.

2004. Folk psychology as mental simulation. In Zalta N. (ed.) *The Stanford encyclopedia of Philosophy* (URL=http://plato.stanford.edu/archives/fall2004/entries/folkpsych-simulation).

2005. Intentional agents like myself. In Hurley S. and Chater N. (eds.) *Perspectives on Imitation* (95-106). Cambridge MA: MIT Press.

2007. Ascent routines for propositional attitudes. *Synthese* 159, 151-65.

Grafton S.T., Arbib M.A., Fadiga L. and Rizzolatti G. 1996. Localization of grasp representation in humans by positron emission tomography. *Experimental Brain Research* 112, 103-11.

Gregory C., Lough S., Stone V., Erzinclioglu S., Martin L., Baron- Cohen S. and Hodges J.R. 2002. Theory of mind in patients with frontal variant frontotemporal dementia and Alzheimer’s disease: theoretical and practical implications. *Brain* 125, 752-64.

Grezes J., Costes N. and Decety J. 1998. Top-down effect of strategy on the perception of human biological motion: A PET investigation. *Cognitive Neuropsychology* 15, 553-82.

Hainline L. 1978. Developmental changes in visual scanning of face and nonface patterns by infants. *Journal of Experimental Child Psychology* 25, 90-115.

Haith M.M., Bergman T. and Moore M.J. 1979. Eye contact and face scanning in early infancy. *Science* 198, 853-5.

Happé F., Malhi G.S. and Checkley S. 2001. Acquired mind-blindness following frontal lobe surgery? A single case study of impaired theory of mind in a patient treated with stereotactic anterior capsulotomy. *Neuropsychologia* 39, 83-90.

Harris P.L. 1992. From simulation to folk psychology: The case for development. *Mind and Language* 7, 120-44.

Haruno M., Wolpert D.M. and Kawato M. 2001. Mosaic model for sensorimotor learning and control. Neural Computation 13, 2201-20.

Hatfield E., Cacioppo J. and Rapson R.L. 1994. *Emotional contagion.* New York: Cambridge University Press.

Hattiangadi J. 2005. The emergence of minds in space and time. In Johnson D. M. and Ernelling C. (eds.) *The Mind as a Scientific Object: Between Brain and Culture (*79–100) New York: Oxford University Press.

Heal J.

1986. Replication and Functionalism. In Butterfield J. (ed.) *Language, Mind and Logic* (135-50) Cambridge: Cambridge University Press.

1995. How to Think About Thinking. In Davies M. and Stone T. (eds.) *Mental Simulation* (33-52) Oxford: Blackwell.

1998. Co-Cognition and Off-Line Simulation: Two Ways of Understanding the Simulation Approach. *Mind and Language* 13, 477-98.

Hempel C.G. 1949. The logical analysis of psychology. In Feigl H. and Sellars W. (eds.) *Readings in Philosophical Analysis* (373-384) New York: Appleton-Century-Crofts.

Hickok G. 2009. Eight Problems for the Mirror Neuron Theory of Action Understanding in Monkeys and Humans. *Journal of Cognitive Neuroscience* 21, 1229-43.

Hobson R.P.

1993. The emotional origins of social understanding. *Philosophical Psychology* 6, 227-49.

2002. *The Cradle of Thought.* London: Macmillan.

Horgan T. and Woodward J. 1985. Folk Psychology is Here to Stay. *Philosophical Review* 94, 197-226.

Horner V. and Whiten A. 2005. Causal knowledge and imitation/emulation switching in chimpanzees (Pan troglodytes) and children (Homo sapiens). *Animal Cognition* 8, 164–181.

Howe M.L. 2000. *The fate of early memories: Developmental science and the retention of childhood experiences*. Washington DC: American Psychological Association.

Hume D. 2003. *A Treatise of Human Nature*. New York: Dover.

Hurley S.L. 2005. Active perception and perceiving action: The shared circuits model. In Gendler T. and Hawthorne J. (eds.) *Perceptual Experience* (205-59) New York: Oxford University Press.

Hurley S.L. 2008. The shared circuits model (SCM): How control, mirroring, and simulation can enable imitation, deliberation, and mindreading. *Behavioral Brain Sciences* 31, 1-58.

Hutto D.D.

1999. A Cause for Concern: Reasons, Causes and Explanations. *Philosophy and Phenomenological Research* 59, 381-401.

2004. The limits of spectatorial folk psychology. *Mind and Language* 19, 548-73.

2006. Four Herculean labours: Reply to Hobson. In Menary R. (ed.) *Radical Enactivism: Focus on the Philosophy of Daniel D. Hutto* (185–221) Amsterdam/Philadelphia: Jon Benjamins.

2007a. *Folk psychological narratives: the socio-cultural basis of understanding reasons*. Cambridge MA: MIT Press.

2007b. The narrative practice hypothesis: Origins and applications of Folk Psychology. *Narrative and Understanding Persons, Royal Institute of Philosophy Supplement* 82, 43-68.

2008a. The Narrative Practice Hypothesis: Clarifications and Implications. *Philosophical Explorations* 11, 175-92.

2008b. Articulating and Understanding the Phenomenological Manifesto. *Abstracta - Linguagem Mente e Ação,* 10-9.

2009. *Folk Psychology as Narrative Practice*. *Journal of Consciousness Studies* 16, 9-39.

Hyslop A. and Jackson F.C. 1972. The Analogical Inference to Other Minds. *American Philosophical Quarterly* 9, 168-76.

Iacoboni M. 2005. Neural mechanisms of imitation. *Current Opinion in Neurobiology* 15, 632-7.

Iacoboni M., Molnar-Szakacs I., Gallese V., Buccino G., Mazziotta J. and Rizzolatti G. 2005. Grasping the intentions of others with one's owns mirror neuron system. *PLOS Biology* 3, 529-35.

Iacoboni M., Woods R.P., Brass M., Bekkering H., Mazziotta J.C. and Rizzolatti G. 1999. Cortical mechanisms of human imitation. *Science* 286, 2526-8.

Iacoboni M., Woods R.P., Brass M., Bekkering H. and Mazziotta J.C. and Rizolatti G. 2001. Reafferent copies of imitated actions in the right superior temporal cortex. *PNAS* 20, 13995-9.

Imamizu H., Kuroda T., Yoshioka T. and Kawato M. 2004. Functional magnetic resonance imaging examination of two modular architectures for switching multiple internal models. *Journal of Cognitive Neuroscience* 24, 1173–81.

Jacob P. and Jeannerod M. 2005. The motor theory of social cognition: a critique. *Trends in Cognitive Sciences* 9, 21-5.

James W. 1890. *The Principles of Psychology*. New York: Dover Publications (reprinted 1950).

Jeannerod M. 1994. The representing brain: Neural correlates of motor intention and imagery. *Behavioral and Brain Sciences* 17, 187-245.

Jeannerod M. and Pacherie E. 2004. Agency, simulation, and self-identification. *Mind and Language* 19, 113-46.

Jenkins J.M. and Astington J.W. 2000. Theory of mind and social behavior: Causal models tested in a longitudinal study. *Merrill-Palmer Quarterly* 37, 369-405.

Happé F. 1995. *Autism: An introduction to psychological theory*. Cambridge: Harvard University Press.

Herman D. 2007. The *Cambridge Companion to Narrative*. Cambridge: Cambridge University Press.

Kalaska J.F., Caminiti R. and Georgopoulos A.P. 1983. Cortical mechanisms related to the direction of two-dimensional arm movements: relations in parietal area 5 and comparison with motor cortex. *Experimental Brain Research* 51, 247-60.

Keysar B. and Bly B*.* 1995*.* Intuitions of the transparency of idioms*:* Can one keep a secret by spilling the beans*? Journal of Memory and Language* 34*,* 89-109.

Keysar B*.,* Lin S. and Barr D.J. 2003. Limits on theory of mind use in adults. *Cognition* 89, 25-41.

Keysers C., Kohler E., Umiltà M.A., Nanetti L., Fogassi L. and Gallese V. 2003. Audiovisual mirror neurons and action recognition. *Experimental Brain Research* 153, 628-36.

Keysers C. and Perrett D.I. 2004. Demystifying social cognition: a Hebbian perspective. *Trends in Cognitive Sciences* 8, 501-7.

Kim J.

1993. *Supervenience and Mind: Selected Philosophical Essays.* Cambridge: Cambridge University Press.

1998. *Mind in a physical world.* Cambridge: MIT Press.

1999. Making Sense of Emergence. *Philosophical Studies* 95, 3-36.

Kinsbourne M. 2004. Imitation as entrainment: Brain mechanisms and social consequences. In Hurley S. and Chater N. (eds.) *Perspectives on Imitation: From Neuroscience to Social Science Volume II* (163-72). Cambridge MA: MIT Press.

Klin A., Jones W., Schultz R., Volkmar F.R. and Cohen D.J. 2002. Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of General Psychiatry* 59, 809-16.

Klin A., Jones W., Schultz R. and Volkmar F. 2003. The enactive mind, or from actions to cognition: lessons from autism. *Philosophical Transactions of the Royal Society B: Biological Sciences* 358, 345–60.

Kohler E., Keysers C., Umiltà M.A., Fogassi L., Gallese V. and Rizzolatti G. 2002. Hearing sounds, understanding actions: action representation in mirror neurons. *Science* 297, 846-8.

Koski L., Wohlschlager A., Bekkering H., Woods R.P., Dubeau M.C., Mazziotta J.C. and Iacoboni M. 2002. Modulation of motor and premotor activity during imitation of target-directed actions. *Cerebral Cortex* 12, 847-55.

Lacquaniti F., Guigon E., Bianchi L., Ferraina S. and Caminiti R. 1995. Representing spatial information for limb movement: role of area 5 in the monkey. *Cerebral Cortex* 5, 391-409.

Leslie A.M.

1982. Discursive representation in infancy. In De Gelder B. (ed.) *Knowledge and representation* (80-93) London: Routledge and Kegan Paul.

1987. Pretense and representation: The origins of ‘theory of mind’. *Psychological Review* 94, 412-26.

1988. Some implications of pretense for mechanisms underlying the child’s theory of mind. In Astington J.W., Harris P.L. and Olson D.R. (eds.) *Developing theories of mind* (64-92) Cambridge: Cambridge University Press.

1991. Theory of mind impairment in autism. In Whiten A. (ed.) *Natural theories of mind: Evolution, development, and simulation of everyday mindreading* (233-251) Cambridge MA: Basil Blackwell.

1994. ToMM, ToBy, and Agency: Core architecture and domain specificity. In Hirschfeld L. and Gelman S. (eds.) *Mapping the Mind: Domain Specificity in Cognition and Culture.* Cambridge: Cambridge University Press.

Leslie A.M. and Frith U. 1988. Autistic children's understanding of seeing, knowing and believing. *British Journal of Developmental Psychology* 6, 315-24.

Lewis D. 1972. Psychophysical and Theoretical Identifications. *Australasian Journal of Philosophy* 50, 249-58.

Lewis M. 1997. The self in self-conscious emotions. In Snodgrass J.G. and Thompson R.L. (eds.) *The self across psychology* (119–42) New York: New York Academy of Sciences.

Lewis M. and Brooks-Gunn J. 1979. *Social Cognition and the Acquisition of Self*. New York: Plenum Press.

Lhermitte F., Pillon B. and Serdaru M. 1986. Human autonomy and the frontal lobes Volume I. *Annals of Neurology* 19*,* 326-34.

Lhermitte F. 1986. Human autonomy and the frontal lobes Volume II. *Annals of Neurology* 19, 335-43.

Lillard A.

1993. Pretend play skills and the child's theory of mind. *Child Development* 64, 348–71.

1997. Other Folk's Theories of Mind and Behavior. *Psychological Science* 8, 268-74.

1998. Ethnopsychologies: Cultural Variations in Theories of Mind. *Psychological Bulletin* 123, 3-32.

2002. Pretend play and cognitive development. In Goswami U. (ed.) *Handbook of cognitive development* (188-205) London: Blackwell.

Lingnau A., Gesierich B. and Caramazza A. 2009. Asymmetric fMRI adaptation reveals no evidence for mirror neurons in humans. *PNAS* 106, 9925-30.

Lipps T. 1903. Einfuhlung, innere Nachahmung und Organempfindung. *Archive fur die Gesamte Psychologie* 1, 185-204.

Locke J.L. 1993. *The child's path to spoken language*. Cambridge MA: Harvard University Press.

MacIntyre A. 1981. *After virtue.* Notre Dame IN: University of Notre Dame Press.

Malatesta C. and Haviland J.M. 1982. Learning display rules: The socialization of emotion expression in infancy. *Child Development* 53, 991-1003.

Martin G.B. and Clark R.D. 1987. Distress crying in neonates: Species and peer specificity. *Developmental Psychology* 18, 3-9.

Maturana H.R. and Varela F.J. *Autopoiesis and Cognition: The Realization of the Living*. Dordrecht: D. Reidel.

Maurer D. 1985. Infants' perception of facedness. In Field T.N. and Fox N. (eds.) *Social perception in infants* (73-100) Norwood NJ: Ablex.

Maurer D. and Barrera M.E. 1981. Infants' perception of natural and distorted arrangements of a schematic face. *Child Development* 52, 196-202.

McGeer V. 2001 Psycho-practice, psycho-theory and the contrastive case of autism: How practices of mind become second-nature. *Journal of Consciousness Studies* 8, 109-32.

McLaughlin B. 1989. Type Epiphenomenalism, Type Dualism, and the Causal Priority of the Physical. *Philosophical Perspectives* 3, 109-135.

Meltzoff A.N.

1988. Infant imitation after a 1-week delay: Long-term memory for novel acts and multiple stimuli. *Developmental Psychology* 24, 470-6.

1995. Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology* 31, 838-50.

2002. Elements of a developmental theory of imitation. In Meltzoff A.N. and Prinz W. (eds.) *The imitative mind: Development, evolution, and brain bases* (19-41) Cambridge: Cambridge University Press.

2004. Imitation and other minds: The “like me” hypothesis. In Hurley S. and Chater N. (eds.) *Perspectives on Imitation: From Neuroscience to Social Science Vol. II* (55-77) Cambridge MA: MIT Press.

Meltzoff A.N. and Brooks R. 2001. ‘Like me’ as a building block for understanding other minds: Bodily acts, attention, and intention. In Malle B.F., Moses L.J. and Baldwin D.A. (eds.) *Intentions and intentionality: foundations of social cognition* (171-91). Cambridge MA: MIT Press.

Meltzoff A.N. and Moore M.K.

1977. Imitation of facial and manual gestures by human neonates. *Science* 198, 75-78.

1983. Newborn infants imitate adult facial gestures. *Child Development* 54, 702-9.

1994. Imitation, memory, and the representation of persons. *Infant Behavior and Development* 17, 83-99.

1995. Infants' understanding of people and things: from body imitation to folk psychology. In Bermúdez J., Marcel A. and Eilan N. (eds.) *Body and the self* (43-69). Cambridge MA: MIT Press.

Menary R. 2006. *Radical Enactivism*. Amsterdam: John Benjamins.

Mill J.S. 1878. An examination of Sir Williams Hamilton's Philosophy 4th Ed. London: Longmans, Green, Reader and Dyer.

Miller P.J., Potts R., Fung H., Hoogstra L. and Mintz J. 1990. Narrative practices and the social construction of self in childhood. *American Ethnologist* 17, 292–311.

Millikan R.G.

1993. *White Queen Psychology and Other Essays for Alice*. Cambridge/London: MIT Press.

2004. *Varieties of Meaning*. Cambridge MA: MIT Press.

Molnar-Szakacs I., Iacoboni M., Koski L. and Mazziotta J.C. 2005. Functional segregation within pars opercularis of the inferior frontal gyrus: evidence from fMRI studies of imitation and action observation. *Cerebral Cortex* 15, 986-94.

Montero B.

1999. The Body Problem. *Nous* 33, 183-200.

2001. Post-Physicalism. *Journal of Consciousness Studies* 8, 61–80.

Morales M., Mundy P. and Rojas J. 1998. Following the direction of gaze and language development in 6-month-olds. *Infant Behavior and Development* 21, 373-7.

Morales M., Mundy P., Delgado C.E.F., Yale M., Messinger D., Neal R. and Schwartz H.K. 2000. Responding to joint attention across the 6 to 24-month age period and early language acquisition. *Journal of Applied Developmental Psychology* 21, 283–98.

Moses L.J. and Flavell J.H. 1990. Inferring false beliefs from actions and reactions. *Child Development* 61, 929-45.

Nelson K.

1996. *Language in cognitive development: emergence of the mediated mind*. Cambridge: Cambridge University Press.

2003. Narrative and the emergence of a consciousness of self. In Fireman G.D., McVay T.E. and Flanagan O.J. (eds). *Narrative and Consciousness* (17-36) Oxford: Oxford University Press.

Nelson K. and Gruendel J. 1981. Generalized event representations: Basic building blocks of cognitive development. In Lamb M. and Brown A. (eds.) *Advances in developmental psychology Volume I* (131-58) Hillsdale NJ: Erlbaum.

Newman-Norlund R.D., Noordzij M.L., Meulenbroek R.G.J. and Bekkering H. 2007. Exploring the brain basis of joint attention: Co-ordination of actions, goals and intentions. *Social Neuroscience* 2, 48-65.

Nickerson R.S.

1999. How we know -and sometimes misjudge- what others know: imputing one's own knowledge to others. *Psychological Bulletin* 125, 737-59.

2001. The Projective Way of Knowing: A Useful Heuristic That Sometimes Misleads. *Current Directions in Psychological Science* 10, 168-72.

Nichols S. and Stich S.

2000. A Cognitive Theory of Pretense. *Cognition* 74, 115-47.

2003. *Mindreading: an integrated account of pretense, self-awareness and understanding of other minds*. Oxford: Oxford University Press.

Ninio A. and Snow C.E. 1996 *Pragmatic Development.* Boulder CO: Westview Press.

Nisbett R.E. and Wilson T.D. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological Review* 84, 231-59.

Norton J.D. 2003. Causation as Folk Science. *Philosophers' Imprint* 3, 1-22.

Oberman L.M., Hubbard E.M., McCleery J.P., Altschuler E.L., Ramachandran V.S. and Pineda J.A. 2005. EEG evidence for mirror neuron dysfunction in autism spectrum disorders. *Cognitive Brain Research* 24, 190-98.

Onishi K.H. and Baillargeon R. 2005. Do 15-month-old infants understand false beliefs? *Science* 308, 255-8.

Perner J.

1991. *Understanding the representational mind*. Cambridge MA: MIT Press

1996. Simulation as explication of prediction-implicit knowledge about the mind: arguments for a simulation-theory mix. In Carruthers P. and Smith P. (eds.) *Theories of Theories of Mind* (90-104) Cambridge: Cambridge University Press.

Perner J., Leekam S. and Wimmer H. 1987. Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology* 5, 125-37.

Perrett D.I. and Emery N.J. 1994. Understanding the intentions of others from visual signals: neurophysiological evidence. *Current Psychology of Cognition* 13, 683-94.

Perrett D., Harries M., Bevan R., Thomas S., Benson P., Mistlin A., Chitty A., Hietanen J. and Ortega J. 1989. Frameworks of analysis for the neural representation of animate objects and actions*. Journal of Experimental Biology* 146, 87-113.

Perrett D., Mistlin A., Harries M. and Chitty A. 1990. Understanding the visual appearance and consequence of hand actions. In Goodale M. (ed.) *Vision and Action: The Control of Grasping* (163–180) Norwood NJ: Ablex.

Pinker S. 1994. *The Language Instinct*. Baltimore: Penguin.

Powers P.S., Schulman R.G., Gleghorn A.A. and Prange M.E. 1987. Perceptual and cognitive abnormalities in bulimia. *American Journal of Psychiatry* 144, 1456-60.

Premack D.G. and Woodruff G. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* 1, 515-26.

Prinz J.J. and Clark A. 2004. Putting Concepts to Work: Some Thoughts for the Twenty-First Century. *Mind and Language* 19, 57-69.

Putnam H. 1990. *Realism with a Human Face*. Cambridge MA: Harvard University Press.

Quine W. 1953. Two dogmas of empericism. In *From a logical point of view* (20–46) Cambridge MA: Harvard University Press.

Ratcliffe M.

2005. Folk Psychology and the Biological Basis of Intersubjectivity. In O'Hear A. (ed.) *Philosophy, Biology and Life* (211-33) Cambridge: Cambridge University Press.

2006. ‘Folk psychology’ is not folk psychology. *Phenomenology and the Cognitive Sciences* 5, 31–52.

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

2008. Farewell to Folk Psychology: A Response to Hutto. *International Journal of Philosophical Studies* 16, 445-51.

Reddy V. 2003. On being the object of attention: implications for self-other consciousness. *Trends in Cognitive Sciences* 7, 397-402.

Reddy V. and Morris P. 2004. Participants Don’t Need Theories: Knowing Mind in Engagement. *Theory and Psychology* 14, 647-65.

Reid T. 1983. Essays on the Intellectual Powers of Man. In Beanblossom R. and Lerher K. (eds.) *Thomas Reid's Inquiry and Essays* (VI, 278-79) Indianapolis: Hacket.

Ricoeur P.

1984. *Time and Narrative*. Chicago: University of Chicago Press.

1992. *Oneself As Another*. Chicago: University of Chicago Press.

Rizzolatti G., Fadiga L., Gallese V. and Fogassi L. 1996. Premotor cortex and the recognition of motor actions. *Cognitive Brain Research* 3, 131-41.

Rizzolatti G., Luppino G. and Matelli M. 1998, The organization of the cortical motor system: New concepts. *Electroencephalography and Clinical Neurophysiology* 106, 283–96.

Rizzolatti G., Fogassi L. and Gallese V. 2000. Cortical mechanisms subserving object grasping and action recognition: A new view on the cortical motor functions. In Gazzaniga M.S. (ed.) *The New Cognitive Neurosciences* (539-52). Cambridge MA: MIT Press.

Rizzolatti G. and Craighero L. 2004. The mirror-neuron system. *Annual Review of Neuroscience* 27, 169–92.

Rochat P. and Hespos S.J. 1997. Differential rooting response by neonates: Evidence for an early sense of self. *Early Development and Parenting* 6, 105-12.

Rollins P.R. and Snow C.E. 1998. Shared attention and grammatical development in typical children and children with autism. *Journal of Child Language* 25, 653-73.

Rorty R.

1979. *Philosophy and the Mirror of Nature*. Princeton: Princeton University Press.

1982. Contemporary Philosophy of Mind. *Synthese* 53, 323-48.

Rosen W.D., Adamson L.B. and Bakeman R. 1992. An experimental investigation of infant social referencing: Mothers' messages and gender differences. *Developmental Psychology* 28, 1172–78.

Roth P.A. 1991. Truth in interpretation: The case of psychoanalysis. *Philosophy of the Social Sciences* 21, 175-95.

Rowe J.B., Toni I., Josephs O., Frackowiak R.S. and Passingham R.E. 2000. The prefrontal cortex: response selection or maintenance within working memory? *Science* 288, 1656-60.

Rowe A.D., Bullock P.R., Polkey C.E. and Morris R.G. 2001. ‘Theory of mind’ impairments and their relationship to executive functioning following frontal lobe excisions. *Brain* 124, 600-16.

Ruby P. and Decety J. 2001. Effect of the subjective perspective taking during simulation of action: a PET investigation of agency. *Nature Neuroscience* 4, 546-50

Russell B.

1940. *An inquiry into meaning and truth*. London: Allen and Unwin.

1948. *Human Knowledge: Its Scope and Limits*. London: Allen and Unwin.

Ryle G. 1949. *The Concept of Mind.* New York: Barnes and Noble.

Sagi A. and Hoffman M. 1976. Empathic Distress in the Newborn. *Developmental Psychology* 12, 175-76.

Sakata H., Takaoka Y., Kawarasaki A. and Shibutani H. 1973. Somatosensory properties of neurons in the superior parietal cortex (area 5) of the rhesus monkey. *Brain Research* 64, 85-102.

Sass L. 1992. *Madness and modernism. Insanity in the light of modern art, literature and thought*. New York: Basic Books.

Saxe R., Carey S. and Kanwisher N. 2004. Understanding Other Minds: Linking Developmental Psychology and Functional Neuroimaging. *Annual Review of Psychology* 55, 87-124.

Scheler M. 1973. *Wesen und Form der Sympathie*. Bern/München: Francke Verlag (Engl. Translation: Scheler M. 1954. The *Nature of Sympathy*. London: Routledge and Kegan Paul)

Scholl B.J. and Tremoulet P.D. 2000. Perceptual causality and animacy. *Trends in Cognitive Science* 4, 299-309.

Sellars W.

1956. Empiricism and the philosophy of mind. *Minnesota Studies in the Philosophy of Science* 1, 253-329.

1963. *Science, perception and reality.* London: Routledge and Kegan Paul.

1975. *Action*, *Knowledge and Reality: Studies in Honor of Wilfrid Sellars*. Indianapolis: The Bobbs-Merrill.

Seltzer B. and Pandya D.N. 1994. Parietal, temporal, and occipital projections to cortex of the superior temporal sulcus in the rhesus monkey: a retrograde tracer study. *Journal of Comparative Neurology* 343, 445-63.

Senju A., Johnson M.H. and Csibra G. 2006. The development and neural basis of referential gaze perception. *Social Neuroscience* 1, 220-34.

Shoemaker S. 1984. *Identity, Cause, and Mind*. Cambridge: Cambridge University Press.

Sleutels J. 1994. *Real knowledge: The problem of content in neural epistemics.* Nijmegen: KU Nijmegen.

Slors M.V.P. 2009. Neural Resonance: Between Simulation and Perception. *Phenomenology and the Cognitive Sciences* 9, 1.

Somerville J. 1989. Making out the Signatures. In Dalgarno M. and Matthews E. (eds.) *The Philosophy of Thomas Reid* (249-273) Dordrecht: Kluwer.

Sperber D. 1993. Interpreting and Explaining Cultural Representations. In Palsson G. (ed.) *Beyond Boundaries* (162-83) Oxford: Berg Publishers.

Sperry R.W. 1969. A Modified Concept of Consciousness. *Psychological Review* 76, 532-6.

Stern D.N. 1985. *The Interpersonal World of the Infant: A View from Psychoanalysis and Developmental Psychology*. New York: Basic Books.

Steuber K.R. 2006*. Rediscovering empathy: agency, folk psychology, and the human sciences.* London: MIT Press.

Stich S. 1983. From Folk Psychology to Cognitive Science. Cambridge, MA: MIT Press

Stich S. and Nichols S.

1992. Folk Psychology: Simulation or Tacit Theory*? Mind and Language 7, 35-71.*

1997. Cognitive Penetrability, Rationality, and Restricted Simulation. *Mind and Language* 12, 297-326.

Stich S. and Ravenscroft I. 1994. What is Folk Psychology? *Cognition* 50, 447-68.

Stone V.E., Baron-Cohen S. and Knight R.T. 1998. Frontal lobe contributions to theory of mind. *Journal of Cognitive Neuroscience* 10, 640-56.

Strawson G.

1999. Self, body, and experience. *Proceedings of the Aristotelian Society (Supplement)* 73, 307–32.

2006. Realistic monism: Why physicalism entails panpsychism. *Journal of Consciousness Studies* 13, 3-31.

Striano T. and Reid V.M. 2006. Social cognition in the first year. *Trends in Cognitive Sciences* 10, 471-76.

Thalberg I. 1964. Emotion and Thought. *American Philosophical Quarterly* 1, 45-55.

Thompson E. 2007. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Cambridge MA: Harvard University Press.

Thompson W.C., Clarke-Stewart K.A. and Lepore S. 1997. What did the janitor do? Suggestive interviewing and the accuracy of children's accounts. *Law and Human Behavior* 21, 405-26.

Tooby J. and Cosmides L. 1995. Foreword to S. Baron-Cohen*.* In *Mindblindness: An essay on autism and theory of mind* (xi-xviii) Cambridge MA: MIT Press.

Tognoli E., Lagarde J., De Guzman G.C. and Kelso J.A.S. 2007. The phi complex as a neuromarker of human social coordination. *PNAS* 104, 8190-5.

Tomasello M.

1988. The role of joint attentional process in early language development. *Language*

Sciences 10, 69-88.

1999. *The Cultural Origins of Human Cognition.* Cambridge: Harvard University Press.

2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Cambridge: Harvard University Press.

2008. *Origins of Human Communication.* Cambridge MA: MIT Press.

Tomasello M., Kruger A.C. and Ratner H.H. 1993. Cultural learning. *Behavioral and Brain Sciences 16*, 495-552.

Tomasello M., Strosberg R. and Akhtar N. 1996. Eighteen-month-old children learn words in non-ostensive contexts. Journal of Child Language 23, 157-76.

Tomasello M., Carpenter M., Call J., Behne T. and Moll H. 2005. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences 28*, 675-91.

Trevarthen C. 1979. Communication and cooperation in early infancy: a description of primary intersubjectivity. In Bullowa M. (ed.) *Before Speech: The beginning of interpersonal communication* (321-47) Cambridge: Cambridge University Press.

Tversky A. and Kahneman D. 1986. Rational choice and the framing of decisions. *Journal of Business* 59, S251-0S278

Umiltà M.A., Kohler E., Gallese V., Fogassi L., Fadiga L., Keysers C. and Rizzolatti G. 2001. I know what you are doing: a neurophysiological study. *Neuron* 32, 91-101.

Van der Geest J.N., Kemner, C., Verbaten, M.N. and Van Engeland H. 2002. Gaze behavior of children with pervasive developmental disorder toward human faces: a fixation time study. *Journal of Child Psychology and Psychiatry* 43, 1-11.

Varela F. 1979. *Principles of biological autonomy*. New York: North Holland.

Varela F., Thompson E. and Rosch E. 1991. *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge: MIT Press.

Vicera S. and Johnson M. 1995. Gaze detection and the cortical processing of faces: evidence from infants and adults. *Visual Cognition* 2, 59-87.

Vinden P.

1996. Junin Quechua Children's Understanding of Mind. *Child Development* 67, 1707-16.

1999. Children's Understanding of Mind and Emotion: A Multi-Culture Study. *Cognition and Emotion* 13, 19-48.

2002. Understanding Minds and Evidence for Belief: A Study of Mofu Children in Cameroon. *International Journal of Behavioral Development* 26, 445-52.

Vogeley K., Bussfeld P., Newen A., Herrmann S., Happé F., Falkai P., Maier W., Shah N.J., Fink G.R. and Zilles K. 2001. Mind reading: neural mechanisms of theory of mind and self-perspective. *NeuroImage* 14, 170–81.

Walker A.S. 1982. Intermodal perception of expressive behaviors by human infants. *Journal of Experimental Child Psychology* 33, 514-35.

Watson J.B. 1913. Psychology as the Behaviorist views it. *Psychological Review* 20, 158-77.

Wegner D.M.

2002. *The Illusion of Conscious Will*. Cambridge: MIT Press.

2003. The mind's best trick: how we experience conscious will. *Trends in Cognitive Sciences* 7, 65-9.

2005. Who is the controller of controlled processes? In Hassin R.R., Uleman J.S. and Bargh J.A. (eds.) *The new unconscious* (19-36) New York: Oxford University Press.

Wellman H.M. 1990*. The Child's Theory of Mind*. Cambridge MA: MIT Press.

Wellman H.M. and Phillips A. 2001. Developing intentional understandings. In Malle B., Moses L.J. and Baldwin D.A. (eds.) *Intentions and intentionality* (125-48) Cambridge MA: MIT Press.

Wellman H.M. and Woolley J.D. 1990. From simple desires to ordinary beliefs: The early development of everyday psychology. *Cognition* 35, 245-75.

Wicker B., Keysers C., Plailly J., Royet J.P., Gallese V. and Rizzolatti G. 2003. Both of us disgusted in my insula: the common neural basis of seeing and feeling disgust. *Neuron* 40, 655-64.

Williams J.H.G., Whiten A., Suddendorf T. and Perrett D.I. 2001. Imitation, mirror neurons and autism. *Neuroscience and Biobehavioral Reviews* 25, 287-95.

Wilson T.D. 2002. *Strangers to Ourselves: Discovering the Adaptive Unconscious*. Cambridge MA: Harvard University Press.

Wimmer H. and Perner J. 1983. Beliefs about Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children’s Understanding of Deception. *Cognition* 13, 103-28.

Wimmer H., Hogrefe J. and Sodian B. 1988. A Second Stage in Children's Conception of Mental Life: Understanding Informational Access as Origins of Knowledge and Belief. In Astington J.W., Harris P.L. and Olson D.R. (eds.) *Developing Theories of Mind* (173-92) Cambridge: Cambridge University Press.

Wittgenstein L. 1953. *Philosophical Investigations.* Oxford: Blackwell.

Wolpert D.M., Ghahramani Z. and Flanagan J.R. 2001. Perspectives and problems in motor learning. *Trends in Cognitive Sciences* 5, 487-94.

Woodward J. 1984. A Theory of Singular Causal Explanation. *Erkenntnis* 21, 231-62.

Woodward A.L.

1998. Infants selectively encode the goal object of an actor's reach. *Cognition* 69, 1-34.

2005. The infant origins of intentional understanding. In Kail R.V. (ed.) *Advances in Child Development and Behavior* (229-62) Oxford: Elsevier.

Youngblade L.M. and Dunn J. 1995. Individual Differences in Young Children's Pretend Play with Mother and Sibling: Links to Relationships and Understanding of Other People's Feelings and Beliefs. *Child Development* 66, 1472-92.

Zahavi D. 2001. Beyond empathy: phenomenological approaches to intersubjectivity. *Journal of Consciousness Studies* 8, 151-67.

Zimmerman A. 2007. The Nature of Belief. *Journal of Consciousness Studies* 14, 61-82.

Index

Index

De menselijke geest in praktijk

Het probleem van intersubjectiviteit

Dit boek gaat over dagelijkse ontmoetingen tussen doodgewone mensen. Of misschien is het beter te stellen dat het gaat over wat er aan deze ontmoetingen *vooraf* gaat, aangezien het probeert inzichtelijk te maken welke praktijken en processes onze interacties met anderen funderen, formeren en faciliteren. Zodoende introduceert het boek een pragmatische visie op *intersubjectiviteit*. Soms gebruiken we het woord 'empathie' om een ervaring van eensgezindheid of verbondenheid te beschrijven die kan ontstaan wanneer we ons soepel door het sociale landschap bewegen. Dit boek, echter, gaat over meer dan empathie in zoverre het afstand neemt van het idee dat we onze één op één relaties met anderen kunnen begrijpen in termen van een unieke modus van bewustzijn. In plaats daarvan benadrukt het dat de mogelijke manieren waarop we ons tot anderen kunnen verhouden niet vooraf gegeven zijn, maar dat deze worden geconditioneerd en gestructureerd door ons lichamelijke bestaan en onze sociale gesitueerdheid.

De meeste hedendaagse benaderingen van intersubjectiviteit kunnen grofweg in twee categorieën worden onderverdeeld: 'theorie theorie' (TT) en 'simulatie theorie' (ST). Theorie theorie beweert dat we in onze ontmoetingen met anderen afhankelijk zijn van een 'volkspsychologische' theorie die ons (tot op zekere hoogte) in staat stelt het gedrag van andere mensen te voorspellen en te verklaren. Sommige voorstanders van TT gaan ervan uit dat zo'n theorie kant en klaar wordt meegeleverd met de geboorte, in de vorm van een geavanceerde biologische module. Het idee is dat pasgeboren en zeer jonge kinderen alleen nog de basisprincipes van deze module kunnen gebruiken, maar dat ze tijdens hun verdere ontwikkeling in toenemende mate leren te exploiteren wat ze eigenlijk al weten (cf. Fodor 1995). Andere bepleiters van TT benadrukken dat de vaardigheid om andermans gedrag te voorspellen en te verklaren niet aangeboren is, maar dat deze zich ontwikkelt op het moment dat kinderen in toenemende mate de wereld gaan verkennen en aan het experimenteren slaan. Volgens deze 'kind-als-wetenschapper' benadering gedragen kinderen zich in hun ontwikkeling net als wetenschappers die vooruitgang boeken in hun onderzoek: ze verkrijgen sociale kennis door nieuw bewijsmateriaal te verzamelen en oude theorieën te (her)interpreteren in het licht daarvan (cf. Gopnik and Meltzoff 1997).

Simulatie theorie (ST) verwerpt het idee dat sociaal begrip een theorie vereist. In plaats daarvan stelt deze theorie dat onze omgang met anderen primair bepaald wordt door ons vermogen om ons in de situatie van anderen te verplaatsen, en ons voor te stellen hoe het zou zijn om ‘in hun schoenen te staan’. Voorstanders van de ‘offline’ ST zijn de mening toegedaan dat zo’n proces wordt aangestuurd door zelfgegenereerde mentale toestanden die worden ingevoerd in ons eigen beslissingsmechanisme en vervolgens worden geprojecteerd op degene die we willen begrijpen of wiens gedrag we willen voorspellen (cf. Goldman 2006). Degenen die een variant van ‘directe simulatie’ verdedigen daarentegen, beweren dat simulatie veel meer is dan alleen een cognitieve heuristiek: het stelt ons in staat om onszelf in de ander te ‘transformeren’, en zo tot sociaal begrip te komen (cf. Gordon 1995).

Ondanks het feit dat TT en ST vaak worden afgeschilderd als bittere rivalen, hebben ze eigenlijk veel gemeen. Om een eerste indruk te krijgen van datgene wat deze posities precies motiveert is het van belang om in te zien dat ze beide een antwoord proberen te geven op een fundamentele vraag over intersubjectiviteit: hoe zijn we überhaupt in staat om te herkennen dat onze medemens ‘begeestigd’ is, net als wijzelf? John Stuart Mill (1878) formuleerde de kwestie als volgt: ‘welk bewijs heb ik, of door welke beschouwingen word ik aangespoord, om aan te nemen dat er andere bewuste creaturen bestaan; dat de wandelende en sprekende gestalten die ik aanschouw en hoor eigen ervaringen hebben en gedachten, of anders gezegd, over een geest beschikken?’ (p.243). Dit probleem staat vandaag de dag bekend als *het probleem van de ‘andere geest’*. Mill was overigens welwillend genoeg om tevens een mogelijke oplossing voor dit vraagstuk aan te dragen: het zogenaamde ‘argument van analogie’. Mill stelde voor dat, aangezien een ieder van ons reeds bekend is met zijn/haar eigen geest en weet hoe deze gerelateerd is aan zijn/haar lichaam, we kunnen afleiden dat dit waarschijnlijk ook wel het geval zal zijn voor de personen om ons heen, op basis van een analogie tussen onze lichamen en die van de personen om ons heen.

Het argument van analogie vormt nog steeds het uitgangspunt voor de meeste versies van ST. Echter, een mogelijk bezwaar tegen Mill’s argument van analogie is het feit dat het gebaseerd is op een inductief argument dat is afgeleid van slechts één enkel geval. Om deze reden benadert TT het probleem van de andere geest vanuit een hele andere hoek.

TT stelt dat we mentale toestanden zoals overtuigingen en verlangens moeten opvatten als theoretische (niet-waarneembare) entiteiten, en beweert dat we deze entiteiten mogen postuleren zolang dit een behoorlijke mate van voorspellende en verklarende kracht met zich mee brengt (cf. Churchland 1988). De specifieke elementen van TT en ST, en de wijze waarop zij het probleem van de andere geest karakteriseren komen uitgebreid aan bod in de eerste hoofdstukken van dit boek.

Het is belangrijk om te realiseren dat TT en ST, door te zoeken naar een antwoord op het probleem van de andere geest, in feite akkoord lijken te gaan met een aantal vooronderstellingen die aan dit probleem ten grondslag liggen. Deze vooronderstellingen zijn van doorslaggevend belang voor de wijze waarop zij intersubjectiviteit karakteriseren:

(i) In de eerste plaats gaan beide posities ervan uit dat onze ontmoetingen met anderen intrinsiek *problematisch* zijn. Het probleem van de andere geest suggereert dat sociale interactie wordt aangestuurd door *twijfel*: hoe kunnen we zekerheid krijgen over het bestaan van de andere geest? TT en ST volgen in de voetstappen van Mill in zoverre ze onze alledaagse omgang met anderen afbeelden als gecompliceerde puzzels, als onzekere expedities naar een verafgelegen en onbekende streek genaamd ‘de andere geest’.

(ii) Echter, het respect dat TT en ST koesteren voor het probleem van de andere geest gaat verder dan een opvatting van sociale interactie als een zoektocht naar zekerheid. Het behelst tevens de impliciete acceptatie van een zekere manier van denken over de geest. TT en ST hanteren een begrip van de geest als een geïsoleerd ‘ik’: een autonome entiteit die enerzijds representatief is voor de buitenwereld en het eigen lichaam, maar er anderzijds ook van afgescheiden is. De geest wordt begrepen als een mysterieuze innerlijke wereld die is afgegrensd van de lichamelijke gedragingen die uiterlijk waarneembaar zijn. Een dergelijk begrip van de geest heeft een rijke geschiedenis, en in hun pogingen om haar oorsprong te achterhalen wijzen filosofen met hun vinger maar al te graag naar Descartes - de grootvader van de moderne filosofie van de geest. Deze beschuldigingen zijn niet geheel onterecht, maar tegelijkertijd moeten we niet vergeten dat voor Descartes het bestaan van de andere geest nog niet problematisch was. Descartes was namelijk in staat om de solipsistische consequenties van zijn methodische twijfel te omzeilen door zich te beroepen op een goedbedoelende God. Maar voor de opvolgers van Descartes, die een theologisch beroep op God niet langer wensbaar achtten, kreeg het spook van het solipsisme steeds duidelijkere vormen. Dit is met name zichtbaar in het werk van die filosofen die tot het Britse empirisme worden gerekend. Het is daarom nauwelijks verrassend dat, voor een filosoof als Mill, het probleem van de andere geest een ‘officieel’ filosofisch probleem wordt.

(iii) Een ander belangrijk idee is het idee dat onze twijfel over de andere geest weggenomen kan worden door middel van een zelfbewuste, methodische en kritische manier van denken. Descartes was de overtuiging toegedaan dat alleen een strikte methode van introspectie kon leiden tot zekere kennis, omdat het de gebruiker een direct besef gaf van de ideeën van de geest. Deze ideeën werden gedragen door een goddelijke autoriteit die ons uiteindelijk verzekerde van het bestaan van de andere geest. Mill daarentegen wenste net zoals zijn tijdgenoten niet langer een beroep te doen op God ter rechtvaardiging van het bestaan van de andere geest. In plaats daarvan trachtte hij het bestaan ervan op radicaal andere wijze te rechtvaardigen. Zijn argument van analogie berust op een *inferentieel* proces dat ons in staat stelt om empirische generalisaties te postuleren met betrekking tot onze mentale toestanden en onze lichamelijke gedragingen, om deze vervolgens toe te schrijven aan anderen op basis van een analogie tussen onze lichamen. Echter, ondanks de enorme verschillen tussen Descartes en Mill, gingen beiden ervan uit dat intersubjectiviteit gekenmerkt wordt door een bewust, cognitief proces - een stapsgewijze procedure die wordt geïnitieerd door een hyperreflexief zelf. Dit idee is nog springlevend in hedendaagse articulaties van TT en ST. Het is veelzeggend dat intersubjectiviteit vandaag de dag nog steeds voornamelijk wordt begrepen in termen van ‘volkspsychologie’: een label dat wordt gebruikt om te benadrukken dat ons alledaags begrijpen van de ander uiteindelijk niets meer is dan een ‘volkse variant’ van de methodische en theoretische benadering die karakteristiek is voor de wetenschappelijke psychologie.

(iv) Ten slotte wordt over het algemeen aangenomen dat ‘denken’ of ‘cognitie’ noodzakelijkerwijs functioneert als *mediator* tussen perceptie en actie. Hurley (2008) noemt dit het ‘sandwichmodel’ van sociale interactie. Volgens dit sandwichmodel is onze omgang met anderen als volgt gestructureerd: uitgangspunt is de observatie van andermans lichamelijk gedrag, maar op dit punt hebben we nog geen hard bewijs voor het bestaan van zijn/haar geest of enig idee van wat er in hem/haar omgaat. Om dit te bereiken moeten we eerst een inferentieel en/of deliberatief proces in werking stellen. Pas als dit proces naar tevredenheid is afgerond, zijn we gereed voor interactie. Het gaat te ver om hier de historische wortels van het sandwichmodel volledig weer te geven. We kunnen volstaan met de vermelding dat zowel Descartes als Mill op eigen wijze aan dit model gecommitteerd waren, net zoals veel hedendaagse versies van TT en ST dat zijn.

De menselijke geest in praktijk

Een van de doelstelling van dit boek is om de bovenstaande karakterisatie van intersubjectiviteit te ondermijnen, en daarmee tevens de verschillende problemen die het met zich mee brengt te lijf te gaan. De meeste benaderingen die het belang van cognitie voor sociale interactie benadrukken, zoals TT en ST, proberen onze kennis van de andere geest te modeleren op de perceptuele capaciteiten van de individuele actor. Dit leidt onvermijdelijk tot wat Dewey (1960) een 'toeschouwer-theorie' van kennis noemde. De pragmatische visie op intersubjectiviteit die ik wil voorstellen, daarentegen, wijst op de *interactieve* in plaats van *perceptuele* natuur van onze kennis van de andere geest. Het woord 'pragmatisch' is afgeleid van het Griekse woord 'pragma', dat zoveel als 'actie' betekent. Maar het ligt ook aan de basis van het woord 'praktijk'. Het uitgangspunt van dit boek is dat intersubjectiviteit mogelijk wordt gemaakt door verschillende interactieve praktijken. Deze praktijken structureren onze ontmoetingen met anderen en vormen een fundament voor sociaal begrip. We zouden kunnen zeggen dat dit boek voornamelijk draait om de *praktijk van de menselijke geest.*

Mijn pragmatische visie op intersubjectiviteit is niet zozeer geënt op één bepaalde theorie, maar tracht eerder een aantal recente inzichten en voorstellen met betrekking tot sociale interactie te integreren en te verenigen. Ze kan worden gezien als een vorm van *enactivisme* voor zoverre ze uitgaat van het aforisme 'weten is doen is zijn'. Noch ons bestaan in deze wereld, noch onze kennis ervan is pre-existent in de zin dat het *vooraf gegeven* is. In plaats daarvan is het 'geactiveerd' - het ontstaat als gevolg van onze interacties met de omgeving en onze medemensen. Het enactivisme benadrukt dat de menselijke geest fundamenteel wordt vormgegeven door ons lichaam (belichaming), en dat deze niet begrepen kan worden in isolatie van onze omgeving (gesitueerdheid). In dit boek ga ik met name in op de vraag hoe het proces van 'geestelijke beschaving' begrepen kan worden in relatie tot onze interacties met *anderen*.

Naast het enactivisme bouwt mijn pragmatische voorstel tevens voort op inzichten van verschillende filosofen uit de fenomenologische en de analytische traditie. Het put uit de fenomenologische traditie (met name het werk van Shaun Gallagher), met als doel de ‘fenomenologie van onzekerheid’ te bevragen die wordt verondersteld door TT en ST, en te demonstreren dat de kern van onze alledaagse omgang met anderen niet exclusief een kwestie van kennis is. Integendeel, veel van wat onze interacties met anderen mogelijk maakt, voltrekt zich *voordat* we er weet van hebben. De fenomenologie van sociale interactie suggereert bovendien dat de expliciete vorm van theoretiseren die wordt verondersteld door TT en ST ‘geen onderdeel is van onze alledaagse praktijk, en niet raakt aan hoe we over onszelf en anderen denken’ (Gallagher 2004, p.202). Mijn voorstel put uit de analytische traditie in zoverre het filosofen als Ludwig Wittgenstein en Wilfred Sellars (en hun hedendaagse representanten zoals Daniel Hutto) navolgt in hun visie op de relatie tussen taal en betekenis. Ik gebruik de inzichten van deze filosofen om kritiek te leveren op de poging om onze kennis van anderen te modeleren op een directe gewaarwording van de eigen geest (ST) of op een theoretisch begrip van psychologische principes (TT). Wat ik van hen overneem is het gegeven dat ons begrip van ‘geest’ en ‘wereld’ niet v*eronderst**eld*** hoeft te worden bij of *constitutief* is voor sociale interactie. Integendeel, ze *komt voort* uit de talige praktijken die onze tweede-persoons interacties structureren. Daarom beroept mijn pragmatische benadering van intersubjectiviteit zich niet op een privé-taal of een set impliciete theoretische regels, maar focust ze in plaats daarvan op bestaande talige praktijken, aangezien deze het in eerste instantie mogelijk maken om een eerste en/of derde-persoons vocabulaire te ontplooien.

Een belangrijke doelstelling van dit boek is om een begrip van intersubjectiviteit te articuleren dat een bredere strekking heeft dan de filosofische term ‘volkspsychologie’. Dat wil niet zeggen dat ik denk dat volkspsychologische processen geen enkele rol spelen in onze omgang met anderen. Maar volgens mij is haar inbreng relatief bescheiden, en is haar functie anders dan over het algemeen wordt aangenomen. De filosofische consensus is dat de volkspsychologie primair betrekking heeft op het genereren van betrouwbare voorspellingen en verklaringen van andermans gedrag. Vaak wordt gedacht dat dit berust op een zeer basale (aangeboren) capaciteit die voornamelijk wordt uitgeoefend in een derde-persoons context – in situaties waarin we niet meer dan toeschouwers zijn, die de handelingen van anderen observeren zonder hierbij tot interactie met deze anderen over te gaan. Dit boek, echter, presenteert een visie op volkspsychologie die stevig geworteld is in een interactieve en tamelijk geavanceerde tweede-persoons *praktijk*.

De grenzen van het pragmatisme

Een de(r)gelijke afbakening van het begrip 'volkspsychologie' maakt dat ik een interpretatie van intersubjectiviteit kan geven die verder gaat dan de gangbare 'mentalistische' variant, en die bovendien ondersteund wordt door bewijsmateriaal van interdisciplinaire aard. Dit boek leunt op de resultaten van verschillende wetenschappelijke disciplines - zoals de ontwikkelingspsychologie en de cognitieve neurowetenschappen - en gebruikt deze ter ondersteuning van de verschillende interactieve praktijken die ze introduceert. Om een voorbeeld te geven: onze ontmoetingen met anderen zijn sterk afhankelijk van zeer basale sensorimotor processen die beschreven kunnen worden in neurobiologische termen. Deze processen maken tot op zekere hoogte een vorm van 'hands-free' intersubjectiviteit mogelijk, en helpen te verklaren waarom veel van onze sociale interacties soepel kunnen verlopen zonder dat er bewuste reflectie bij aan te pas hoeft te komen.

Dit leidt onvermijdelijk tot vragen over de status van empirische resultaten in het debat over intersubjectiviteit. Hoewel er in het soort pragmatisme dat ik in dit boek verkondig veel aandacht is voor wetenschappelijk bewijsmateriaal, betekent dit niet automatisch dat ik een vorm van reductionisme of instrumentalisme wil verdedigen. Het is eerder zo dat ik intersubjectiviteit wil bestuderen vanuit de vraag wat mensen *doen* om anderen te begrijpen. Het pragmatisme dat ik in gedachten heb richt zich op actuele tweede-persoons praktijken. Het vraagt: hoe kunnen we beschrijven wat hier gebeurt? En: wat ligt daaraan ten grondslag? De eerste vraag betreft de fenomenologie van intersubjectiviteit. Om haar op de juiste wijze te beantwoorden, moeten we een beroep doen op wat Gallagher en Brøsted Sørensen (2006) 'front-loaded phenomenology' noemen: we moeten een goede beschrijving van onze alledaagse interacties met anderen geven, die dan vervolgens weer kan dienen als input voor wetenschappelijk onderzoek. De tweede vraag suggereert dat we verschillende problematische kwesties rondom intersubjectiviteit kunnen verhelderen door onderzoek te doen naar de manier waarop intersubjectiviteit zich ontwikkelt, en wat de randvoorwaarden voor deze ontwikkeling zijn.

Ik realiseer me maar al te goed dat het pragmatisme als filosofisch programma haar grenzen kent. Het soort pragmatisme dat ik in dit boek verdedig is echter zeer bescheiden. Het bouwt voort op en geeft verdieping aan een interessante gedachtegang van Goldman (1989), die opmerkt dat filosofische verklaringen van intersubjectiviteit slechts dan behulpzaam kunnen zijn, wanneer ze verenigbaar zijn met wetenschappelijke inzichten en met een juiste karakterisatie van wat mensen concreet doen wanneer ze anderen trachten te begrijpen.

Een overzicht van dit boek

In de eerste twee hoofdstukken van dit boek wordt ingegaan op de zogenaamde ‘interne’ problematiek van TT en ST, dat wil zeggen, de problemen die ontstaan wanneer we akkoord gaat met een bepaalde interpretatie van intersubjectiviteit. Er worden zowel conceptuele als fenomenologische argumenten gegeven om aan te tonen dat zowel TT als ST een zeer armoedige en problematische verklaring van sociale interactie voorstaan. Deze hoofdstukken bevatten ook een kritische beoordeling van het empirische bewijsmateriaal waaraan beide partijen appelleren ter onderbouwing van hun claims. Dit varieert van onderzoek in de ontwikkelingspsychologie (b.v. resultaten van de false-belief test) tot experimenten in de neurobiologie (b.v. de vondst van ‘spiegelneuronen’).

Het is echter ook mogelijk om TT en ST verklaringen van intersubjectiviteit op een fundamenteler niveau te bevragen. Zo’n meer hermeneutisch-georiënteerde analyse stelt ons in staat om te onderzoeken tot op welke hoogte deze posities worden geïnspireerd door gemeenschappelijke opvattingen over intersubjectiviteit. Deze vooronderstellingen worden besproken en bekritiseerd in het derde hoofdstuk.

De pragmatische visie die ik wil voorstellen in dit boek situeert intersubjectiviteit in een tweede-persoons interactieve praktijk. Het beoogt een verdere articulatie van Gallagher’s voorstel (zie bv. Gallagher 2005) dat een brede reeks van belichaamde praktijken ons verschillende aangeboren en vroeg ontwikkelende capaciteiten laten ontplooien die ons van een basaal sociaal begrip voorzien - wat Trevarthen (1979) als ‘primaire intersubjectiviteit’ bestempelde. In de loop van de ontwikkeling worden deze capaciteiten meer en meer ingebed in een bredere sociale en pragmatische context, en dit stelt ons in staat om te participeren in praktijken met een gezamenlijke focus (zogenaamde ‘secundaire intersubjectiviteit’). Dit is het onderwerp van hoofdstuk vier.

Belichaamde en gesitueerde praktijken staan niet op zichzelf. Integendeel, ze zijn afhankelijk van en gestroomlijnd door onze lijfelijke bestaan, en bouwen voort op ervaringen die voortkomen uit het hebben van een lichaam met verschillende sensori-motor gestuurde capaciteiten. Hoofdstuk vier laat tevens zien hoe gecompliceerde neurobiologische processen ons niet alleen kunnen voorzien van een minimale vorm van zelfbewustzijn, maar ook van een primitief besef van anderen. Hoewel belichaamde en gesitueerde praktijken de 'base-line' voor sociaal begrip constitueren, betekent dit zeker niet dat ze de mogelijkheden voor intersubjectiviteit uitputten. De ontwikkeling van taal en de opkomst van andere vaardigheiden (zoals temporele integratie, (auto)biografisch geheugen en perspectief nemen) stellen ons in staat te participeren in *narratieve praktijken*, waardoor we leren om ons begrip van zelf en ander verder te verfijnen en verdiepen (Hutto 2007, Gallagher and Hutto 2008). Dit wordt besproken in het eerste deel van hoofdstuk vijf.

Narratieve praktijken kunnen ook verklaren hoe we onze intrede maken in wat Sellars de 'redelijke ruimte' noemde, en de vaardigheid verwerven om de handelingen van anderen te interpreteren in termen van *redenen*. Het tweede deel van hoofdstuk vijf bespreekt de sterke en zwakke punten van Hutto's (2007) 'narrative practice hypothesis', volgens welke kinderen de kunst van de volkspsychologie meester worden door directe interactie met specifieke (volkspsychologische) verhalen die gaan over wat het betekent om redelijk te handelen. Ik stel voor dat kinderen in eerste instantie alleen in staat zijn om de handelingen van anderen te interpreteren tegen de achtergrond van een factieve, gedeelde wereld. Echter, de acquisitie van mentale concepten leidt tot een enorme verbetering van hun interpretatieve vaardigheden, aangezien het hen in staat stelt om de redenen van anderen te *individualiseren* op een wijze die aansluit bij hun psychologische opmaak.

Dit alles leidt uiteindelijk tot een benadering van intersubjectiviteit die het probleem van de andere geest in sterke mate *trivialiseert.* Daarbij verwerp ik de vier bovengenoemde ideeën over intersubjectiviteit die TT en ST er op na houden, en argumenteer in plaats daarvan dat: (i) we niet moeten instemmen met de opvatting dat sociale interactie per definitie problematisch is, (ii) de notie van de menselijke geest die aan deze opvatting ten grondslag ligt fundamenteel verkeerd is, (iii) onze alledaagse sociale ontmoetingen niet per se afhankelijk zijn van theoretische interventies, aangezien (iv) ze stevig gegrondvest zijn in tweede-persoons interacties die kunnen worden begrepen in directe actie-perceptie koppelingen.

Curriculum vitae

Leon de Bruin was born in Nijkerk on June 15th, 1979. He earned his BSc degree in clinical psychology in 2003 at the University of Utrecht, and graduated cum laude from the Free University of Amsterdam in 2004 with a MA degree in philosophy. From 2005 until 2009, he worked as a PhD student at the University of Leiden on the problem of the other mind from an interdisciplinary perspective. In addition to scientific research, he taught several courses in philosophy. In 2007, he completed his MA thesis (cum laude) in cognitive neuropsychology. Since September 2009 he works as a Post-doctoral Research Fellow at the University of Bochum.