universität wien

DISSERTATION

Titel der Dissertation

What is Folk Psychology and Who Cares?

The debate between simulation theory and theory theory from the perspective of the philosophy of mind

Verfasser

M.A. John Michael

angestrebter akademischer Grad

Doktor der Philosophie (Dr.phil.)

Wien, 29. September 2009

Studienkennzahl lt. Studienblatt:	A 092 296
Dissertationsgebiet lt. Studienblatt:	Philosophie
Betreuerin / Betreuer:	Ao. Univ.-Prof. Mag. Dr. Elisabeth Nemeth/ Univ.-Prof. Dr. Friedrich Stadler

Contents

Part I: An introduction to and assessment of the debate between simulation theory and theory theory from the perspective of the philosophy of mind

<u>Part II: Improving on simulation theory in light of recent empirical findings</u>

Part I: An introduction to and assessment of the debate between simulation theory and theory theory from the perspective of the philosophy of mind

Chapter 1:
Introduction

1.1 What is folk psychology?

There has been a lively debate in philosophy of mind as well as in psychology, primatology and the neurosciences over the past 30 years or so about the nature of so-called folk psychology. The central controversy centers around the issue of how to account for our everyday, commonsense psychology. How is it that we can effortlessly predict and understand other people's behavior in most everyday situations? If I see my colleague hiding behind a statue of Gustav Mahler while our mutual boss walks by on the way to a meeting that we should all be going to, I understand that he does not want to go to the meeting, that he does not want our boss to see him because our boss might then order him to go the meeting, and that hiding behind a statue will prevent our boss from becoming aware of his presence because our boss cannot see through statues. The tactic makes sense to me. Similarly, if a woman says to her husband, "I am leaving you." and he responds "Who is he?", the woman knows perfectly well what inferences he has drawn and what his response has to do with her statement, although on the face of it the one has nothing to do with the other. What these cases are intended to show is that making sense of other people's behavior requires us to go beyond their *physical* behavior and ascribe *mental* states and processes to them.

Philosophers like to summarize all the intervening mental states and processes by talking about beliefs and desires. Beliefs and desires are alike in that they represent possible states of the world. But they differ with respect to what one calls, following Searle (1983), direction of fit. Beliefs have a mind-to-world direction of fit because they are true if the way they represent the world matches the way the world objectively is, whereas desires have a world-to-mind direction of fit because they are fulfilled if the world comes to match the way in which they represent it. In other words, beliefs aim to match the world, whereas desires aim to make the world match them. Taken together, a person's beliefs and desires can be said to give an adequate, if abstract, explanation of their behavior, since people act to get what they desire in the way that is most efficient given what they believe about the world. For this reason, folk psychology is sometimes also called "belief-desire psychology".

In fact, folk psychology has a number of different names. Some philosophers, who resist the theoretical flavor of the mainstream conception of folk psychology, prefer to speak of "commonsense psychology" or "everyday psychology". But, on the other hand, some like the term "folk psychology" for just the same reason, since it suggests that our everyday psychological practices are more like a folk art or craft, such as basketweaving or black magic, than they are like a science, being made up of implicit skills and rules of thumb rather than explicit, general laws. I am sympathetic to the skepticism about the term "folk psychology", since the term implies something more abstruse than the simple, everyday understanding of others' behavior that we want to account for. But I do not think "everyday psychology" is a good substitute, since people, perhaps surprisingly, tend to take it to mean something like everyday psychoanalysis, or explanations of people's behavior in terms of hidden motivations and indirect strategies – e.g. "He spoils his son because he is afraid he will otherwise not love him", "She takes care of her sick aunt so that people will think she is a good person". This sort of case is also interesting and forms a part of what we want to explain, but primarily we are concerned with cases that are so simple that people do not think they require explanations at all: they are just plain common sense. Of course, what we want to account for is how it is that these cases appear not to require explanations, even though they patently do require some sort of thought process that takes into account other people's thought processes. But since it is the unproblematic, obvious nature of everyday, commonsense psychological understanding that is at the center of interest here, I personally prefer the term "commonsense psychology". Nevertheless, when I am presenting other people's ideas, I will tend to use their terminology, and since "folk psychology" is by far the most established term, it will be the term that I primarily use. At any rate, I will try to be explicit about what is meant by "folk psychology", "commonsense psychology", "everyday psychology" and other terms when they come up.

In fact, the list of possible names is still not complete. I will also have to mention at least one other popular name for folk psychology: psychologists like to refer to our "theory of mind". This term goes back to a classical paper by Premack and Woodruff (1978) that initiated a wave of research in primatology and psychology, especially developmental psychology, about how best to account for folk psychology or theory of mind. Premack and Woodruff define a theory of mind as follows:

> In saying that an individual has a theory of mind we mean that the individual 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 (1978, p.515).

Although Premack and Woodruff, with their talk of unobservable states, clearly think of a theory of mind as really being something like a theory, it is important to stress that the word "theory" in the term "theory of mind" does not commit one to this view. I know this sounds strange, but it really is considered an open question[1]. This openness is revealed by the fact that one of the theories psychologists consider as potentially accounting for our everyday psychological practices is called the theory theory, which is the theory that theory of mind is a theory. This obviously implies that theory of mind might not be a theory at all, despite the name. You might wonder what else it could possibly be then. Well, that is precisely the question I will focussing on. The debate that has been in full swing since the 1980s between theory theory (TT) and its main rival, simulation theory (ST), will be my central topic. Let me briefly characterize each.

TT, which has been prominently supported by (among others) Alison Gopnik, Henry Wellmann, Josef Perner, Simon Baron-Cohen and Alan Leslie is a view that was originally developed by Wilfred Sellars and David Lewis, and was the default conception of folk psychology until its rival, ST, arose. According to TT, folk psychology is a largely unconscious theory including theoretical entities (mental states such as beliefs and desires) and general psychological laws linking them to each other and to perception and behavior. There are important differences between the different versions of TT. The most important distinction is between what I will call the empiricist account (a.k.a. child-scientist-theory) and what I will call the nativist account (a.k.a. modularist theory). The empiricists think that children acquire ToM in a way that can profitably be compared to scientific history, i.e. to the way in which scientists gather data, formulate and test hypotheses, develop a theory, refine it, replace it, etc. The nativists, on the other hand, think that folk psychology springs from an innate module that simply matures according to its own timetable rather than having to be learned at all. They still think the term "theory" is useful, but (by analogy to Chomsky's linguistics) take it to to refer to representations of entities and laws that are structured like a theory although they are not consciously accessible and are utilized automatically rather than deliberately.

According to ST, on the other hand, we put ourselves into others' shoes and know somehow directly, on the basis of our own experiences, how we would act or think or feel,

[1] Occasionally people add qualifications to signal this openness: e.g. "theory of mind broadly construed" (Tomasello 2005).

and then expect the same of others. The "radical" version espoused by Robert Gordon denies that either psychological laws or mental concepts are generally employed in folk psychological prediction (Gordon 1986, 1995). One of the other main advocates of ST, Alvin Goldman, accepts that mental concepts play a role in simulations, insofar as one must identify someone's mental states in order to set the parameters of a simulation, and also insofar as one must identify and categorize one's own mental state upon completing the simulation procedure before one can ascribe it to another person (1995). But he differs from the theory-theorists in that he claims that there is an introspective component to mental concepts that cannot be captured by their functional roles. Moreover, in deriving a prediction of someone's action from their input states, one employs the same procedures that the target person employs (e.g. practical reasoning), and not separate folk psychological inferential procedures (Goldman 1995, 2006). The third main proponent of ST that I will be discussing, Jane Heal (Heal 1986, 1995), is most interested in the role of simulation in understanding others' behavior *as rational*. She asserts that there is no theory of rationality that could replace our own intuitions about what is rational when it comes to interpreting others' behavior as rational. Despite these differences, there is a central insight that is common to the various simulationist accounts, namely that *we undergo some of the same processes in explaining and predicting others' behavior as we undergo when we ourselves are deciding upon, planning and executing actions, and that this overlap obviates the need for extra, specifically folk psychological knowledge.*

The empirical work on theory of mind that has underlain and structured the debate in psychology, neuroscience, primatology and, increasingly, philosophy got started with the aforementioned paper by Premack and Woodruff (1978). Quickly, the discussion became more specific, and this was due in part to the interventions of philosophers. In the experiments reported by Premack and Woodruff, a chimp appropriately selects a picture of another actor's next move (piling up crates to get at a dangling banana), supposedly demonstrating understanding of the actor's goal (i.e. a mental state). In their commentaries on Premack and Woodruff, Dennett (1978) and Harman (1978) point out a shortcoming of the tests. It could be that the chimp herself would pile the crates up and is just selecting the picture that matches the way she would act without any understanding of or interest in how the other would act. The same would go for beliefs as for goals or desires: if one chimp acts as though she knows that another chimp knows where the banana is, this could simply be because she herself knows where the banana is and does not even concieve of the possibility that another actor may be ignorant of or misinformed about this fact. Obviously, this criticism touches upon the central issue of the TT/ST debate: to what extent and when do we (or chimps) apply mental concepts to understand others' action as opposed to just using ourselves as a model and expecting the same behavior from others without needing to apply any mental concepts or even being aware of what we are doing? Dennett suggested a way of testing whether chimps were "merely" simulating or actually applying mental concepts, which was taken up Perner and Wimmer (1983) and has been the centerpiece of theory of mind research ever since – the famous false-belief task.

In a false belief task, the subject has to predict another actor's behavior given that the other actor has a false (mistaken) belief. So, for example, in the original Perner and Wimmer test, the subject (a child, in this case) sees a series of cartoon pictures. In the pictures, a puppet watches as a chocolate bar is placed in a drawer. The puppet leaves the room, and the chocolate is switched to the cabinet. The puppet then returns to the room to retrieve the chocolate. Given that the puppet was not present during the switch, it will presumably have a false belief that the chocolate is in the drawer and therefore look in the drawer. 3 year-old children were found to be poor at this test, mistakenly expecting the puppet to look in the cabinet, where they themselves would look. In contrast, the 5 year-olds tested were quite proficient.

Coming up with the right answer requires the subject to make a prediction that would differ from how she herself would act. This is supposed to rule out the possibility that they are using their own knowledge rather than a representation of the other's knowledge. Success at a false belief task was therefore supposed originally to indicate the presence of a theory of mind – conceived really in a robust sense as a theory. But when ST arose a few years later (circa 1986), simulation theorists turned the argument on its head, pointing to young children's difficulties with the false belief task as evidence that young children were using a simulation strategy. Exactly what change takes place between 3 and 5 that enables children to succeed at the task was and is controversial: theory theorists have the option of admitting that children simulate until they acquite a proper theory of mind, or arguing that even young children employ a theoretical approach – but it just happens to be an inadequate one. Simulation theorists could either argue that children simulate until 5 and then begin to use something like a theory (Goldman 2006), but only selectively in special cases (such as false belief tasks), or that they always simulate, but between 3 and 5 get better at performing more complex simulations as would be demanded by a false belief task (Gordon 1986, 1995). Hence, what was supposed to be a sort of crucial experiment did not serve to resolve any issues so much as to initiate an intense and complex debate that is still ongoing.

1.2 Who cares?

At first glance it is not clear why this debate should be of interest to philosophers of mind. After all, why should people's common-sense notions about the mind be relevant to a philosophical analysis of what the mind is, of how we know it, etc.? And yet lots of philosophers are interested in the folk psychology debate, i.e. the debate about the different versions of ST and TT and the various combinations thereof.[2] My aim in chapter 2 will be to to make this intelligible by showing how the developments in philosophy of mind over the past fifty years led up to the present discussion. I sketch the development of the folk psychology debate in the philosophy of mind from Ryle and Sellars up to the present. I attempt to show that the interest in folk psychology reflected a natural way to think about the mind-body problem subsquent to the linguistic turn: to leave ontological issues aside and investigate the way in which mental concepts function in explanations of behavior, how they relate to other concepts, whether they are reducible, eliminable, etc.

I argue that a further shift (which I call the empirical turn) began in the 1980s and continues today: the assumption that our everyday psychological competence is best accounted for by postulating a folk psychology featuring mental concepts and psychological generalizations is increasingly viewed as problematic. One important cause of this shift was the appearance of an alternative account of folk psychology, namely ST, at which point the hitherto unquestioned account came to be known as TT. But – although the history of the folk psychology discussion is to a great extent the prehistory of TT, ST also has a philosophy-historical background, and, as is only fair, I will turn my attention to this background in the second part of chapter 2.

Numerous commentators have already pointed out the similarity between ST and various hermeneutic predecessors, from Vico to Dilthey to Collingwood, as well as the relationship between ST and the discussion of empathy in psychology about 100 years ago (e.g. Theodor Lipps) (Blackburn 1995, Gordon 1995, Tomasello 1999, Goldman 2006, Stueber 2006). There is therefore no need for me to deal with this material at length, so I will merely touch on a few themes that will be of immediate interest to me. I will, however, spend

[2] There is of course a veritable flood of literature on folk psychology; the interested reader is referred first and foremost to the following anthologies, which contain classical papers on theory-theory and simulation theory: Carruthers P. and Smith P. (eds) 1996; Davies M. and Stone T. (eds) 1995a; Davies, M. & Stone, T. (eds) 1995b.

a bit more time discussing a different historical background, which has not been dealt with at all so far in connection with ST. People sometimes refer to Piaget and to Quine as two predecessors to ST, but nowhere have I seen mention of any connetion between the two. As it happens, I think there is evidence for a systematic and historical connection between the two that is based upon the tradition of psychophysical parallelism initiated by Gustav Theodor Fechner. Although this tradition, as I have said, is not the only or even in any sense the main historical predecessor to ST, it does have certain advantages as a historical backdrop within this discussion. Apart from actually being historically more directly linked to the present discussion (as I will demonstrate), it also has the advantage of having primarily to do with the mind-body problem, which is not true of Dilthey or Lipps. Moreover, it can be and has been seen as a main source of the same functionalist paradigm in which TT belongs. In fact, parallelism is historically a sort of common source for both TT and ST. For this reason, as we will see, a brief look at its history will be quite useful to us in investigating how TT and ST can be compared, contrasted and/or combined, and how their relevance to the mind-body problem can be articulated.

Apart from giving a historical explanation of how we came to the present situation, I will seek to justify the historical developments I recount by showing how philosophy stands to gain by looking at empirical work concerning the skills and practices that make up folk psychology, and concerning the sort of meta-reflection and justifications actually given in everyday life. I will argue that empirical work on folk psychology reveals that our pre-reflective commonsense understanding of the mind does not commit us to the way of understanding mental concepts that has framed the discussion of the mind-body problem in analytic philosophy so far. Indeed, it is an open question just what construal of mental concepts best suits our everyday understanding. And I think that it is worth investigating whether the set of issues making up the mind-body problem perhaps do not arise for our real everyday understanding of the mind, or whether they are perhaps more easily tractable with respect to this understanding of the mind than with respect to an understanding of the mind that results from a priori philosophical analysis. In short, the working hypothesis I propose is that the mind-body problem can be simplified by linking it up with empirical work in psychology if we frame it as a set of issues surrounding everyday mental concepts, which are best addressed with the combined efforts of empirical psychology, philosophy of mind and language, and philosophy of science. Of course, many philosophers are likely to be skeptical of such an approach, arguing that philosophy does not want to reconcile natural science with everyday conceptions of the mind but with the traditional philosophical conception of the mind. And I must admit that my approach runs the risk of simply changing the subject and leaving the traditional mind-body problem untouched. In order to avoid this danger, I will have to try to change the subject in a satisfying way: i.e. to demonstrate persuasively that we can get satisfying answers to the traditional philosophical questions by starting from an empirical theory about our everyday conception of the mind. But this is a task that I will work at gradually and will not be able to complete within the limited scope of this dissertation. Anyway, I do not have any illusion that the whole mind-body problem can be knocked out in one fell swoop; in the best case, I hope to indicate a productive direction for further research.

1.3 Overview

After completing the historical preliminaries just referred to (chapter 2), I will set out the most prominent versions of TT (chapter 3) and ST (chapter 4), and discuss the most relevant strengths and weaknesses of them. I will then (chapter 5) discuss attempts to decide between them, to combine them, or to throw them both out and start fresh. My conclusion is that there is no uniform distinction to be made between the two competing theories; rather, the various versions represent a broad range of ideas, many of which may prove useful in ongoing empirical research. But instead of throwing both out, I suggest trying to make use of the distinction, and suggest ways of doing so. I also point out that the mutual interdependence of TT and ST is to be expected if one takes seriously the analogy to scientific practice, since simulations are generally developed, utilized and interpreted in combination with various theoretical elements in scientific practice. I then attempt to make use of the analogy to science in developing a hybrid theory by comparing the development of simulation systems in science to cognitive development in children, during which one's own mind and body are developed as a system with which to simulate other members of one's culture.

The upshot of my comparison of ST and TT is that ST does not succeed as a replacement of TT, but that the insight it introduces (i.e.that in making sense of others' behavior *we undergo the same procedures that we would undergo if we ourselves were deciding upon, planning or executing an action in the same circumstances*) should be developed further and specified more clearly in ongoing research, ideally in in a way that reflects and also heuristically benefits ongoing research. Chapters 6, 7 and 8 undertake to contribute to this effort. Chapter 6 is devoted to interpretations of mirror neuron (MN) research (Gallese, Rizolatti, Fadiga, Keysers, Oberman, Pineda, Goldman, Jacob, Csibra), since this work constitutes support for ST insofar as ST would predict that resources for action (e.g. the motor system) would be used in action understanding. But since action understanding appears to require a more abstract kind of representation than motor representation (since one action can be carried out with different movements and different actions can be carried out with one and the same movement in different contexts) and to incorporate contextual information, the role of mirror neurons is likely to be contingent upon their integration with other areas (e.g. STS, SMC).

I suggest that the contribution of MNs to action undertstanding can be grapsed theoretically by expanding the concept of simulation to include simulations of *one's own* past or imagined perceptual experiences along the lines of embodied (a.k.a. perceptual or simulationist) theories of concepts (Barsalou 1999, 2003, Prinz 2002, Lakoff and Gallese 2004). Chapter 7 is therefore devoted to a discussion of theories of concepts, and attempts to locate TT and ST as special cases within broader theories of concepts. Embodied, or simulationist – theories of concepts offer a way to show how action concepts (as well as other concepts) could be represented in a more abstract way than the motor or perceptual systems alone could, but without introducing amodal symbols in the sense of a language of thought. Extending this account to mental concepts, as would be necessary at least for some forms of action understanding, I turn in chapter 8 to theories of action planning and action understanding that prominently invoke the term simulation. Research on covert or simulated action (Jeannerod, Decety, Pacherie) also offers interesting perspectives upon how agent-neutral representations are combined with other factors to distinguish between one's own actions and those of others, suggesting a unified account of mental concepts for first- and third-person ascriptions while still leaving room for privileged access to one's own intentions at least in most cases. I also look at recent work on metacognition, which offers one more helpful ingredient to a simulatinoist theory of mental concepts. Metacognition research presents a deflationary way to conceive of introspection and can legitimately serve as a model for the comparatively direct access that ST thinks we have with respect to our own minds.

I therefore return to metacognition in the speculative closing reflections I offer in chapter 9, where, after summing up the main points of the dissertation, I sketch a philosophical approach to mental concepts that reflects the work discussed up to that point. Metacognition plays a key role in that approach, but since metacognition, introspection, and other simulationist strategies can only be useful in understanding others insofar as we are similar to others, I note that our metacognitive or introspective access to our own minds needs to be linked up with a functionalist account of our epistemic access to minds in general.

In spelling out this idea, I return to the analogy to scientific practice (introduced in chapter 5) invited by the names of both theories (ST and TT) and thereby draw upon the resources of philosophy of science. Tomasello's (see especially Tomasello 1999) findings concerning the importance of social cognition (from gaze following to joint attention, joint action, cooperation, etc.) in cognitive development suggests the following idea: folk psychological practices of ascribing mental states to others constitute a self-validating system. On the one hand, their reliability depends on others' similarity to us, just as a simulation system is reliable insofar as it is similar to the target system. On the other hand, children's use of these practices enables them to learn language and other uniquely human cognitive traits imitatively from adults, and thereby to become more similar to other members of their culture, just as simulation systems are constantly updated with new theoretical knowledge about the target systems, by trial and error, by technological advances, etc. The result of this inscreasing similarity to others during cognitive development is that children's folk psychological skills improve (since these depend on similarity to others), which makes them more able to learn from others and also makes it easier for *others* to interpret *them* using folk pscyhology. This account gives a partial explanation of why our folk or commonsense psychology works. It can help to link functionalist elements of theories of mental concepts with simulationist elements such as metacognition our theory of mental concepts, and is therefore the centerpeice of the concluding sketch of a hybrid theory of mental concepts.

Chapter 2
A Bit of Historical Background

2.0 Introduction

In this chapter, I would like to provide some background to the folk psycholog debate from within the recent history of philosophy. This is intended to make the interest that philosophers have recently been taking in folk psychology intelligible, and (relatedly) to introduce some of the central systematic issues that are at stake in the folk psychology debate.

2.1 Folk psychology and the mind-body problem (A prehistory of TT)

2.1.1 From Ryle to Lewis: The Ramseyfication of the mind

The discussion about folk psychology in the philosophy of mind can be traced back as far as you please, of course, but a reasonable place to locate the beginning of the current debate is Ryle's (Ryle 1949) logical behaviorism, since it initiates the post-linguistic-turn formulation of the mind-body problem, which amounts to casting the mind-body problem as a set of issues surrounding folk psychology. Instead of primarily asking ontological questions about what the mind is, or even epistemological questions about how we have knowledge of it, we ask how our mentalist language works, whether it is indispensable, and how it relates to other uses of language. Ryle's core idea is that mental concepts do not refer to hidden states and processes in a mysterious medium called the mind, but are ways of talking about dispositions to behavior. On this view, folk psychology is not in any danger of being replaced by neuroscience, since the criteria for ascribing mental states are behavioral and contextual, and therefore not within the province of neuroscience.

A further reason why folk psychological explanations, for Ryle, are not in competition with neuroscientific explanations is that they are *not causal but conceptual*. This is a very important point. Folk psychological explanations make sense of people's behavior by embedding it in a context of reasons; they make it appear rational. It is at least prima facie plausible to say that this distinguishes them from explanations in natural science, which embed events in contexts of causes[3]. Scientific explanations succeed when they show that an event occurred because some other event(s) occurred. Saying that the bridge collapsed because there was an earthquake can succeed as an explanation because weh know that bridges often collpase when there are earthquakes, and/or because there is a causal chain that

[3] If you are hostile to causes, think of statistical regularity.

could be spelled out in terms of vibrations underground, gravity, the structural features of the bridge, etc. Saying that Gustav bought a new hat because he has a date tonight cannot succeed on the same terms. There is probably no robust regularity between having dates and buying hats[4], and it is far from clear that anyone could fill in the details of a causal chain. But the explanation works anyway because it makes the purchase appear rational against a background of beliefs and desires, and we regard others as rational.

The classical presentation that was to shape the current folk psychology discussion, though, was conceived as a departure from Ryle and is due to Wilfred Sellars (Sellars 1956). Sellars' view, which takes folk psychology to be a sort of proto-scientific theory of mind, profoundly shaped the discussion within mainstream philosophy of mind until the 1980's. Now, under the label theory-theory, it is still one of the leading candidates. The main innovation vis-à-vis Ryle is the idea that when we ascribe mental states to others to explain and/or predict their behavior, what we are doing is postulating theoretical entities much in the way that scientists postulate theoretical entities to help explain and predict observable phenomena. Sellars puts his idea in terms of a myth featuring our fictitious ancestor Jones. I will quote him at some length, since this passage constitutes the starting point for the entire folk psychology discussion:

> In the attempt to account for the fact that his fellow men behave intelligently not only when their conduct is threaded on a string of overt verbal episodes – that is to say, as we would put it, when they "think out loud" – but also when no detectable verbal output is present, Jones develops a *theory* according to which overt utterances are but the culmination of a process which begins with certain inner episodes. *And let us suppose that his model for these episodes* which initiate the events which culminate in overt verbal behavior *is that of overt verbal behavior itself. In other words, using the language of the model, the theory is to the effect that overt verbal behavior is the culmination of a process which begins with "inner speech*" (Sellars 1956, pp. 317-318).

An important aspect of this proposal, of course, is that the theoretical inner states that are postulated as causes of overt behavior are conceived as linguistic. The advantage of this is that sentences lend themselves quite readily to being used in explanations, especially when the propositional contents of those sentences contain objects and events in the world that figure in the behavior to be explained. So, if the aim is to explain some overt behavior, say Heinrich's boarding the U1 at Schwedenplatz in the direction of Reumannplatz, it is obviously explanatorily relevant to say that this behavior is caused by a set of beliefs and

[4] Not to mention the difficulty that the date does not precede the purchase. This can be dealt with by substituting "making a date" for the date itself, but it highlights the teleological character of folk psychological explanations.

desires that includes “Heinrich desires an Eismarillenknödel”, “Heinrich believes that Tichy has the best Eismarillenknödel”, “Heinrich believes that Tichy is at Reumannplatz”, etc. Specifically, pointing to such unobservable theoretical entites as these has the effect of making the behavior appear to be *rational*. This is a central feature of folk psychological explanations. To see why, just imagine that you have ascribed this set of beliefs to Heinrich and then see him exit the U1 at some stop before Reumannplatz. You will have to revise your understanding of his behavior. Either you can conclude that he is irrational, or you can speculate that your ascription of a set of beliefs and desires was not accurate – i.e. either one of them is wrong or there is some other relevant belief you don’t know about. Clearly, we are much more inclined to choose this latter option in most cases in real life. That is because we want to make sense of other people’s actions, and this means making their actions seem to be rational.

This rough-and-ready apparatus for prediction and explanation of actions by appeal to beliefs and desires (and other intentional states) can be readily complemented with additional theoretical resources – e.g. for inferring new beliefs from standing beleifs (e.g. If Sam believes that gambling towns are full of crooks and that Las Vegas is a gambling town, he will believe that Las Vegas is full of crooks), for deriving beliefs form peceptions (e.g. If Frank sees the cat on the mat he will form the belief that there is a cat on the mat).

If the very notion that we employ a theory of mind, even tacitly, in everyday life, seems too ridiculous to you to entertain (if that is so, don’t worry; a lot of people have this reaction), then it might help to bear in mind the epistemological backdrop against which Sellars proposed this myth. He was attacking what he called the “myth of the given”, namely the notion that some of our beliefs are infallible because the facts that make them true are “given” to us experientially. The myth of the given provides that, although I may be mistaken in my belief that I am currently seeing a red ball, I cannot be mistaken in my belief that I am currently having the experience of seeing a red ball, since this latter belief pertains not to the world but to my sense-data, which are directly given to me[5].

There is a powerful intuition here, and one should try to do justice to it even if one does not want to retain the concept of sense-data or of an infallible source of empirical knowledge. On Sellars’ analysis, which is a highlight of twentieth century philosophy, the grain of truth to this intuition is that for a normal perceiver under normal conditions, seeming to see red simply means seeing red, since the decisive criterion for something’s being red is that normal perceivers seem to see red when they look at it under normal conditions. But

[5] The echoes of Descartes are too obvious to necessitate discussion.

before they can make use of this source of knowedge, they have to learn that they are normal perceivers. They have to observe other people seeing red things and saying “That is red”, and also being applauded by others when they themselves look at red things and say “That is red.” So, what is “given” to them prior to or independently of this learning process is not any knowledge about anything at all - neither about the redness of things nor about their experience of redness. It is simply a perceptual experience that they learn to conceptualize as pertaining to redness by correlating it with verbal reports of redness. Once they have achieved this, it is true that knowledge of the redness of objects is given to them whenever they seem to see red objects under normal conditions, but the ease of this transition belies a complex process of concept acquisition within a linguistic community.

Getting back to Sellars’ theory of mind myth, the point is simply to show that we can make sense of the idea that we refer to mental states to explain behavior – that is, we are not, like Ryle, limited to descriptions of physical events - without falling back into Cartesianism. We do not need to think that we have a different kind of epistemological access to these mental states than we have to physical events. Although the whole idea about Jones inventing folk psychology is not to be taken seriously, it demonstrates that mental states need not be of some special substance in order to be explanatorily useful; they could have the same *epistemological* status as theoretical entities.

Now, this only means that we do not *need* to posit a special epistemological mode of access. We still might want to propose something of the sort, but it would be based upon this more fundamental public negotiation of mental concepts that Sellars describes, and would thus avoid Cartesianism. We can accept that we know our minds more direclty than we know others’ minds but assert that this directness belies a complex history just as the direct access to knowledge about the redness of objects on the basis of seeming to see red belies a complex learning history. In this case, the process would entail that we learn to conceptualize internal states and processes by correlating them with verbal reports made by others in our linguistic community. This is an important point, because one problematic aspect of functionalist accounts of mental concepts such as Sellars’ is that they do not seem to leave room for privileged access to one’s own mind. And yet surely we do not ascribe ourselves beliefs by theorizing about what we have perceived and what inferences we might have drawn. We just know what we believe or desire, right?. Here is what Sellars says about this:

> [O]nce our fictitious ancestor, Jones, has developed the theory that overt verbal behavior is the expression of thoughts, and taught his compartriots to make use of the theory in interpreting each other’s behavior, it is but a short step to the use of this

> language in self-description. Thus, when Tom, watching Dick, has behavioral evidence which warrants the use of the sentence (in the language of the theory) “Dick is thinking ‘p’”… Dick, using the same behavioral evidence, can say, in the language of the theory, “I am thinking ‘p’”. … And now it turns out – need it have? – that Dick can be trained to give reasonably reliable self-descriptions, using the language of the theory, without having to observe his overt behavior. Jones brings this about, roughly, by applauding utterances by Dick of “I am thinking that p” when the behavioral evidence strongly supports the theoretical statement “Dick is thinking that p; and by frowning on utterances of “I am thinking that p,” when the evidence does not support this theoretical statement. Our ancestors begin to speak ot 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* (Sellars 1956, p. 320).

Sellars’ account was developed a bit further by David Lewis in a famous paper in 1972 (Lewsi 1972), in which he proposes a method of defining mental concepts functionally:

> Collect all the platitudes … regarding the causal relations of mental states, sensory stimuli, and motor responses. … Add also all the platitudes to the effect that one mental state falls under another … Perhaps there are platitudes of other forms as well. Include only the platitudes which are common knowledge amongst us: everyone knows them, everyone knows that everyone else knows them, and so on. (Lewis 1972, p. 256.)

An important feature of this approach, which is the forerunner of today’s theory-theory, is that mental concepts are essentially linked to psychological laws, since they are defined by their nomological relations to each other and to perceptions and behavior. One speaks here of “Ramsey sentences” (after the philosopher Frank P. Ramsey) for defining theoretical terms, e.g.:

> (The belief that it is raining = x)
> ∃x (the perception that is raining causes x, and x together with the desire to stay dry causes the behavior of carrying an umbrella)

One important achievement of the Sellars/Lewis theory is to re-introduce the causal element to folk psychological explanation while retaining the rationalist element. They do so by conceiving of the normative aspect of rationality in terms of the normative aspect of function. Beliefs and desires do not just tend to have particular effects; they have the function of doing so and are indeed defined in terms of these effects. Notice the similarity between explanations in terms of function and explanations in terms of rationality. In both types of case, if we characterize the initial conditions in a situation and make a prediciton, but if the expected effect does not occur, we will be inclined to revise our characterization of the initial conditions. With functional explanations, this is only partially true, since we are prepared to accept that functions are sometimes unfulfilled. With rationality, it is a bit more extreme. If we ascribe a constellation of beliefs and desires to someone and predict an action on the basis of it, but they do not perform the action, we are more inclined to revise the ascription of beliefs and desires than to regard the prediction as having failed. So, if I ascribe to Julio the desire to go to the movie and the belief that the D train goes to the cinema, but he does not board the train, I will probably conclude that he did not have that desire after all, that he had a stronger desire, or that he did not know the D train goes to the theater. In short, I will revise the initial conditions until Julio's action makes sense. The connections between beliefs and desires, on the one hand, and actions on the other, do not seem to be subject to empirical scrutiny. They seem to have the character of a priori principles.

Donald Davidson (1970) famously concluded from this sort of consideration that psychology does not admit of laws and that the relation between the mental and the physical is therefore anomolous: physical explanations appeal to causes, psychological explanations appeal to reasons, and never the twain shall meet. There are ways to reply to this. One could recall, for example, that predictions in natural science also involve loads of ceteris paribus conditions, and thus leave room for revising and ad hoc adjustments without throwing out a hypothesis. But even if this does relativize Davidson's contrast between the mental and the physical, there is still a contrast if not, as Davidson would have it, an unbridgeable gulf.

A functionally conceived folk psycholgical theory, then, does indeed yield pretty good explanations and is practically useful. On the other hand, it is not clear how these mental states, processes and events could relate to physical states, processes and events, and therefore whether and how folk psychology could relate to scientific psychology. Functional states have the advantage that they could be irreducible and ineliminable without implying dualism. They could be irreducible because they simply pick out different classes of referents than physical terms do. For example, it could be that the belief p plays a specific role in cognitive processes in humans (and perhaps also in computers) but is realized by different neural or silicon) states. It would, in this sense, be multiply realizable. If this is so, we would not be able to substitute a physically defined term for the functionally defined belief p, since p could correspond to any number of physical states. In other words, mental state types would not be identical with physical state types. Nevertheless, it may be the case that each token mental state is identical with a token physical state. So there is no need to appeal to dualism. On this account, it follows that mental concepts could be ineliminable as well, namely if the explanations of behavior one gives in terms of them are the best explanations available. So, the detour via an anlysis of mental concepts did indeed yield an attractive position with respect to the mind-body problem. But functionalism also has certain difficulties.

2.1.2 Eliminativism, intentional realism and the intentional stance

In the discussion concerning the place of mental conepts in a naturalistic worldview, there are a few classical issues that everyone more or less consensually regards as explananda. These problematic features of mental concepts make them appear incompatible with a naturalistic worldview, and thereby give rise to a conflict between our putative everyday understanding of the mind, and scientific psychology – in other words, the mind-body problem. Obviously, I do not have the space to discuss these issues in any detail, but it is worth mentioning the most some central points for the sake of contextualizing the discussion of folk psychology. The most salient potentially problematic features of mental concepts are (drum roll, please…):

a) Intentionality: the mental states postulated by the theory have the peculiar feature of referring to objects or states of affairs in the world (or even to non-existing things), and being differentiated from each other in part on the basis of their content. But there is no clear physicalist way to explain the relation obtaining between a mental state and something in the world that it is about.

b) Semantic Opacity: Talking about mental states creates opaque contexts in which co-referential terms are not substitutable. If we ascribe to someone the belief that Mark Twain was a good author, we cannot infer that s/he believes that Samuel Clemens was a good author, since s/he might not know that Samuel Clemens and Mark Twain refer to the same person.

c) Privileged Access: In the first-person case, it seems highly implausible that we ascribe mental states to ourselves in a purely theoretic manner, observing our behavior and postulating theoretical entities to account for it. Rather, we seem to have some sort of privileged access to our own mental states. But how could this be if the states in question are theoretical postulates? Functionalism seems not to have an answer on hand. On the other hand, if we do have some sort of privileged access to them, then how? If it is to be via internal perception, then it is questionable how they could ever be scientifically relevant or useful, since they could only be experienced from the first-person perspective.

d) Mental Causation: If mental states are postulated as causes of behavior, then we run into trouble with the principle of causal closure of the physical world. Most philosophers agree that naturalism in even the weak sense of a worldview that is compatible with the best available natural sciences demands acceptance of the principle that physical events have only physical causes. So, insofar as behavior is physical, it should have physical causes. If mental events, then, are to be causes of behavior, then they would have to be identical with or instantiated by physical events. But then the explanatory work is being done by the physical states, so it is unclear why we should need to use mental terms at all.

e) Qualia: In contrast to other natural phenomena (at least prima facie), mental states involve an experiential or qualitative aspect that is given only to the person having them, i.e. they involve qualia. This aspect appears not to be captured by a functional definition. In fact, it is arguable that qualia could be inexplicable from a scientific perspective, which would make them a challenge to naturalism.

Functionalism certainly does not lose the game because it has particular problems with qualia and privileged access. But we should at least be able to demand of a theory of folk psychology that it show *that* these features arise even if it cannot reconcile them with natural science, and functionalism does not do this. Still, every theory has strengths and weaknesses, and should be allowed a bit of ad hoc refining here and there, so this criticism is no knock-out argument against functionalism. At the moment, I only want to point out where the weaknesses are.

At any rate, if folk psychology is committed to mental states that exhibit strange features that make them hard or impossible to naturalize, then a theory of folk psychology should reflect that. It would then be an open and legitimate question whether folk psychology could or should eventually be improved upon or eliminated from our worldview. This is the background assumption made by eliminativists (Churchland 1981), who argue that mental state terms are animistic remnants and scientific psychology will eventually develop more precise concepts and abandon mental concepts just like physics eventually threw out ether. But there are options for countering eliminativism. Let me just mention one realist and one instrumentalist alternative in order to the give the flavor of the philosophical discussion in the 1980s, when the TT/ST debate got underway.

One popular realist way to counter the threat of eliminativism, exemplified by Jerry Fodor (1975), is to characterize mental states analogously to the machine states of a computer. A mental state such as a belief is a configuration of symbols. These symbols play a role in producing behavior solely in virtue of their syntactical relations to other symbols. Its intentional content cannot be causally relevant as such, but only insofar as it is mirrored by the functional role of this symbol within the mind. But note: this theory is perfectly compatible with these mental symbols not being experienced as having intentional content. Thus, it cannot distinguish between the intentionality of human minds on the one hand, and computer programs or thermometers on the other. And it is compatible with all this computation not being experienced by anyone at all.

Attempts have also been made to invoke adaptive function in order to introduce the semantic content that is necessary for intentionality. Standing in a reliable causal relation with a property of the environment is not sufficient for intentionality, otherwise sunburn would have the sun as its intentional content. to be one that stands in a reliable causal relation to some other state (outside the mind). Since our brains evolved to track relevant features in the environment, we can say that they have this function. Dretske (1981) and Millikan (1984) take an intentional state to be one that has the adaptive function of being reliably caused by particular external conditions. But this is still really a causal dependency and not a semantic one. There are plenty of physiological mechanisms with adaptive functions but without intentionality. Such a functional relation is in fact no different from that linking a thermometer to the information it displays.

Another influential approach, this one instrumentalist in character, is not to try at all to reconcile folk psychology with naturalism at all, but to regard the postulates of folk psychology as useful fictions. On this view, espoused by Dennett (1978, 1981), we are justified in ascribing mental states to people on the basis of their usefulness in predicting behavior. But when we do so, we are not really postulating any objects that would have to be squeezed into a naturalistic framework. Dennett distinguishes three stances that one can take in predicting a system's behavior:

1) In the physical stance one regards the system as a physical system and uses the laws of physics.
2) In the functional stance one regards the system as having a function or purpose and predicts its behavior accordingly.
3) In the intentional stance, one ascribes intentional states such as wishes and desires to a system and predicts its behavior accordingly, i.e. one practices folk psychology.

A main point of criticism that has always been brought forth against Dennett is that he cannot explain why folk psychology is so useful if the objects it supposedly invokes in its explanations are mere fictions. For Dennett, folk psychology works well because we are products of evolution. Although the brain is a syntactic machine, it has evolved in such a way that it mimics a rational agent with adaptive beliefs and desires. So folk psychology is a theory that explains behavior by appealing to its adaptive function. Two problems with Dennett's view are:

1) He cannot really distinguish between the functional and the intentional stance. A thermometer or a computer are the products of an engineer just as we are the products of evolution. So why do we ascribe beliefs and desires to people but not to them?
2) Since he accounts for the usefulness of folk psychology being appealing to adaptive function, Dennett can only account for its usefulness in the case of adaptive behavior. This seems strange in light of the fact that there was surely folk psychology before evolutionary theory. But, in any case, it also seems that we can understand and predict others' behavior not only when it is adaptive, but also sometimes in the case of maladaptive behavior such as smoking. In fact, it is more natural to suppose that folk psychology works whenever we can bring our own experiences to bear upon interpreting others, which is the case wherever we have similar experiences, either because of our common descent or because of our common cultural background.

Aside from arguing that folk psychology is indeed empirically well-founded, another response to eliminativism is to question whether the folk psychological apparatus they want to eliminate really exists in the first place. Or, more moderately, to claim that eliminativism starts out from an inaccurate picture of folk psychology. In a certain sense, this is what ST claims. It is similar to Dennett's position in that it denies that our real, everyday commonsense understanding of the mind commits us to the view that people's behavior really is caused by functionally defined beliefs and desires[6], but it differs in that it is in a certain sense more extreme: it denies that we even make such ascriptions even on an instrumental basis. Folk psychology, according to ST, works in a different way altogether, namely by expecting others to act as we would act, regardless of what the hidden causes of the relevant behavior are. Hence, ST pulls the rug out from under the discussion of eliminativism. There will be plenty of time later on to say what ST replaces this picture with; right now I want to step back and provide a bit of historical background to ST.

[6] As we shall see, at least one version of ST, namely Goldman's, does have it that we ascribe beliefs and desires, but denies that they are functionally defined or in any sense theoretical terms. More on this later of course.

2.2 Locating ST within the history of philosophy

2.2.1 Rationality and the first-person perspective

Now, obviously, if we can predict and understand people's behavior by putting ourselves in their shoes, and if this is to provide an alternative to the view that we understand others' behavior theoretically, then we must have some special sort of understanding of our own minds or at least of our own actions (privilged access), which can be projected onto others because others are similar to us in a way that inanimate objects are not[7]. Whether this understanding of our own actions means introspection of our mental states as causing our actions is a separate question, and it is a controversial one at that.

Although the way in which this idea is currently discussed and employed in philosophy of mind and related empirical disciplines is novel, the basic idea is not entirely new. At least as far back as Vico and Herder, there has been a hermenutic tradition according to which our approach to understanding human actions and their products in different historical settings can and should draw from our similarity to and/or historical relationship with the actors in question. Famously, Dilthey and others took this as grounds for thinking that psychology should be methodologically distinct from natural science. Since this background is commonly referred to in discussions of ST, and since, as I have eluded to, I am interested in looking at a different historical background instead, I am not going to talk about hermeneutics or the *Verstehen/Ekläaren*-debate associated with Dilthey and co. here (Blackburn 1995, Gordon 1995, Tomasello 1999, Goldman 2006, Stueber 2006). Suffice it to say that these figures also target the rationality of human action as a feature that is potentially distinct from other phenomena in nature. This, as we have also seen, is also true of everyone I mentioned in the discussion of the historical background of TT, although some emphasize it more than others, and different people have drawn different conclusions from it.

The hermeneutic tradition is a bit different, though, in one crucial respect, which I would like to mention, since ST is also different in this same respect from TT. Specifically, they think that the way in which we generally understand others' rationality cannot be captured by a theory. Note that this does not yet exclude the possibility that a theory could be developed that would do a better job. It only means that the way in which we ordinarily do so is not best characterized by a theory or by analogy to a theory. Of course, some, such as Dilthey (and Jane Heal, as we shall see) deny that our everyday intuitive theory of rationality

[7] Other organisms may be somewhere in between. People from exotic cultures and young children may also be a bit more difficult to understand by simulation. In short, there seems to be a continuum.

could be replaced, or that rationality is the kind of thing that ever could be captured by a theory, but that is a further point that I will leave to the side for the time being.

There are two thinkers who are often referred to as predecessors of ST, but never in conjunction with each other, but who both focus on the just this theme of being able to construe others' behavior as rational because we ourselves are rational: Quine and Piaget. The appeal to Quine (e.g. Goldman 2006) is based on occasional remarks about the principles of charity that he thinks we must apply in order to even begin trying to interpret other people's utterances. The situation is clearest in the case of radical translation, wherere the interpreter does not know the language of the natives he is studying and must begin by assuming they are rational and that they will tend to have the same intuitions as he does about what stimuli are irelevant in a situation. Applying these principles, according to Quine, involves something like doing a simulation of the other person's perspective, for example:

> Practical psychology is what sustains our radical translator all along the way, and the method of this psychology is empathy: he imagines himself in the native's situation as best he can (Quine 1990, 46).

With Piaget, the connection with ST is generally (e.g. Perner 1991) based upon Piaget's (1959) concept of childhood egocentricity, according to which children initially treat their own spatial perspective as the only perspective. The famous experiments show that they attribute to others the same visual informaiton that they themselves have, and only gradually overcome this tendency by learning to take others' perspectives. Although the analogy is limited, since Piaget is talking specifically about spatial perspective and not about cognitive states such as beliefs and desires in this context, there is a clear parallel to ST here in that children's access to their own mental states are treated as relatively unproblematic, and others' minds are apprehended by projecting this first-person awareness onto them.

What ties Quine and Piaget together (historically and systematically) is a philosophical tradition in which the point I have been making about about rationality emerges from a discussion of the mind-body problem and the nature of mental concepts – the tradition of psychophysical parallelism. Before I begin my brief historical excursion on the history of parallelism, let me just whet your appetite with a quote from Piaget that brings out the importance of this principle to Piaget. The quote, which is from the *Genetic Epistemology*, published in 1950, introduces a section entitled 'Le Parallélisme Psycho-physiologique':

> It is now the moment to investigate the scope of the famous principle of parallelism, upon which the weight of all the difficulties rests that are proper to the genetic mode of

> explanation and perhaps to psychology in its entirety[8].

2.2.2 The History of Parallelism

I am going to start by introducing the principle of parallelism. That will involve a first section in which I present G.T. Fechner's parallelism, and a second section in which I set out the modified parallelism espoused by Ewald Hering (and with which Ernst Mach is also associated). This is the version that Piaget draws upon. In both of these first two sections, I will also characterize Fechner's and Hering's ideas on self-organizing systems, since the novel interpretation Piaget gives of parallelism involves re-routing Fechner's theory of self-organization in a direction already suggested by Hering. In the third section, I will discuss Piaget's reception of parallelism (his understanding of the tradition and his sources) and explain the novel interpretation he gives of parallelism. I will be arguing that Piaget's interpretation is a sensible modification of parallelism that deals well with at least one point of criticism brought forth against earlier forms of parallelism. We will see that the prominent role Piaget ascribes to physiology in explaining development undermines the prevalent view that his account of cognitive development is intellectualist, solipsistic or disembodied. I will also briefly discuss the function that parallelism serves within the broader context of Piaget's thought in general, particularly with respect to his genetic epistemology. He says that parallelism provides the ultimate justification of his genetic epistemology by 'closing the circle of the sciences.' Interestingly, Quine picked up on just this aspect of Piaget's work. We will see that Piaget's interest in parallelism provides corroboration for Michael Heidelberger's hypothesis of historical and systematic links between Fechner and Quine/Davidson.

2.2.2.1 Fechner's Parallelism

Psychophysical parallelism is a term that goes back to the work of the philosopher and founder of psychophysics, Gustav Theodor Fechner. According to Fechner's (1860) most basic, empirical formulation, parallelism is a heuristic principle according to which one should be able to find a physical correlation for every mental event. He writes:

> The most general law is this: that nothing in the mind exists, arises or ends without something in the body existing, arising or ending along with it[9].

[8] Piaget 1950, 161 my translation ('Le moment est donc venu d'examiner la portée du principe fameux du «parallélisme», qui supporte en fait le poids de toutes les difficultés propres à l'explication génétique et peut-être de la psychologie toute entière').

[9] Fechner 1861, 211, my translation ('Das allgemeinste Gesetz ist dieses: daß nichts im Geiste bestehen, entstehen, gehen kann, ohne daß etwas im Körper mit besteht, entsteht, geht)'.

This formulation does not seek to explain the correlation, but regards it as a functional relation and does not address the issue of causality. The principle is therefore a working hypothesis, or empirical postulate, to be adopted by empirical psychology.

Fechner also formulated a version of parallelism that seeks to interpret the principle by denying any causal relation between the mental and the physical and espousing a dual-aspect theory, which explains the difference between the mental and the physical as resulting from a difference between an inner and an outer perspective:

> Insofar as the difference in appearance is due to the difference in the standpoint of observation…the same essence has two sides, a mental, or psychical, one when it appears to itself, and a material, or bodily, one when it appears to someone else in another form, but body and mind, or body and soul, are not two fundamentally different essences that are attached to each other[10].

I have a unique perspective upon my own mental events insofar as I experience them from the inside, but they can also be studied from the outside, i.e. scientifically, as physiological events. Since the inner perspective cannot be captured or replaced from an objective external perspective, it is irreducible.

But, and here we come to a point of *criticism* against parallelism, all causal relations obtain between physical states that are externally observable, and not between mental events and physical events or among mental events, so the irreducible inner perspective contributes only a phenomenal aspect and not additional information about the causes that lead to behavior. Hence, it can safely be disregarded by an empirical psychology that seeks to explain the causes of behavior. For this reason, many philosophers and psychologists, such as William James and Alfred Binet, criticized Fechner's interpretation of parallelism, lamenting that it led to an unsatisfactory *epiphenomenalism.* This criticism has two aspects. First, the inner, or first-person, perspective appears not to grasp any causal relations and is thus *not necessary* for explanations of psychological processes. Secondly, it appears to be subjective and therefore *not legitimate* for a scientific psychology.

Aside from these and other points of criticism, parallelism was also undermined by the damage Fechner seems to have done to his own reputation by espousing panpsychism later in his career. He thinks that, just as we can ascribe an inner perspective inductively to other

[10] Fechner 1851, 321, my translation ('Die Verschiedenheit der Erscheinung hängt an der Verschiedenheit des Standpunkts der Betrachtung…In sofern hat dasselbe Wesen zwei Seiten, eine geistige, psychische, sofern es sich selbst, eine materielle, leiblich, sofern es einem anderen als sich selbst in anderer Form zu erschienen vermag, nicht aber haften Körper und Geist oder Leib und Seele als zwei grundwesentlich verschiedene Wesen an einander').

humans (by analogy to ourselves), we can also ascribe inner perspectives even to non-living systems and to the universe as a whole. It is worth noting, however, that this panpsychist position has a reasonable core to it. Rather than seeing Fechner as liberally postulating weird spiritual properties, one can take this position as an attempt to understand mental properties in such a way that they are continuous with other properties in nature. So Fechner links freedom of the will with anomolous or spontaneous kinds of motion, which he claims also exist in nature and which ground his indeterminism. For Fechner, though, this spontaneous motion is not very common in nature, since there is also a tendency toward stability on the part of all systems as well as the universe as a whole. Freedom of the will is a relict of an earlier stage of the universe during which it was less stable and more spontaneous, anomalous. Hence his panpsychism is a naturalist account of mental properties.

2.2.2.2 Hering's Psychophysiological Parallelism

Be that as it may, the need to distinguish oneself from Fechner's panpsychism may have been one reason why some supporters of parallelism did not explicitly call themselves parallelists. Limiting parallelism to physiology would exclude inner lives for other physical entities, such as stones and the universe as a whole. Ernst Mach and Ewald Hering (1834-1918) are two cases in point. Both sought a way to employ Johannes Müller's notion of specific nerve energies to distinguish physiologically between different sensation qualities, and resisted Helmholtz's idea that unconscious inferences, i.e. psychological processes, are responsible for interpreting intrinsically homogenous sensations. Mach wrote in a notebook entry of 1896:

> I believe that the thought expressed here, according to which just so many different kinds of physiochemical processes are suspected as there are sensation qualities to be distinguished, likewise has heuristic value, and that this thought can hope for some support from the physiological-chemical side…If I am not mistaken, only Hering still upholds Müller's original teaching[11].

Hering pursued a physiological explanation of Fechner's psychphysical law, which specified a proportional relation between stimulus intensity and logarithm stimulus strength, where stimulus strength is measured in just-noticeable differences in intensity. Hering expected to

[11] Mach E. In: R. Haller and F. Stadler, (eds).(1988): 190, my translation ('Ich glaube nun daß der hier ausgesprochene Gedanke, nach welchem so vielerlei physikalisch-chemische Nervenprocesse zu vermuten sind, als man Empfindungsqualitäten zu unterscheiden vermag, ebenfalls heuristischen Wert hat, und daß derselbe hoffen kann, einmal von physiologisch-chemischer Seite gefördert zu warden… Wenn ich nicht irre, halt nur Hering allein die ursprüngliche Müllersche Lehre noch aufrecht').

find physiological mechanisms to explain this relationship, and also criticized Fechner for generalizing the logarithmic relation to sensation as a whole, suspecting that it could turn out to be linear in most cases. Hering writes:

> …my conception of the functional relationship between body and soul stands in better agreement with the philosophy of Fechner than his own psychophysical law[12].

Hence, the term psycho*physiological* parallelism presumably refers back to this physiological interpretation of Fechner's psychophysical law and psychophysical parallelism – which, especially in the hands of the Gestalt psychologists, also came to be called 'Isomorphism'. Incidentally, Piaget uses the terms isomorphism and parallelism interchangeably.

In 1909, Wolfgang Köhler took up an idea suggested by Mach in the *Analyse der Empfindungen* about how to investigate the role of the tympanic membrane in hearing, especially in the fixation of tones. He had 2 mirrors placed on his eardrum. A steady stream of light was directed at one mirror and reflected via the second mirror back to a recording apparatus. He took the results to support a 'physiological interpretation of Fechner's psychophysical law'; specifically, his interpretation of the results was that the tensor tympani is 'an accommodation muscle which tenses the eardrum more strongly the greater the sound intensity, just as the corresponding muscle regulates pupil width in the eye'. He goes on to say that the tensor tympani inhibits the vibration of the eardrum with increasing intensity 'in just such a way that a logarithmic function appears'.

A year later, Köhler repeated a study Helmholtz had done, in which subjects are asked to associate the tones of thirty different tuning forks with vowels. Interestingly, the tones that subjects associated with pure vowel sounds ranging from German u (engl: oo) to German i (engl: ee) formed a series of ascending octaves. Köhler rejected Helmholtz's interpretation of this phenomenon, according to which psychological processes (i.e. unconscious inferences) interpret the physical tones, which themselves possess only frequency and amplitude and can therefore be described in terms of pitch and loudness. Köhler denies that these uninterpreted physical tones are present at all, asserting that the "vowel qualities are not qualities that the tonal region has alongside others"; rather, they are "the only qualities it has at all." In fact, knowledge of a tone's vowel character is necessary to produce a judgment of pitch.

Hering wanted to push physiology as far as possible in explaining things like spatial perception. Hence he avoided postulating obscure psychological processes along the lines of

[12] Hering 1876, 310, my translation (…meine Auffassung des functionellen Zusammenhanges zwischen Leib und Seele [steht] mit der Philosophie Fechner's in besserem Einklange als sein eigenes psychophysisches Gesetz.).

Helmholtz's unconscious inferences. But this goes hand in hand with his insistence on the autonomy of physiology vis-à-vis physics, physiology being a science of living matter, whereas physiology for Helmholtz is just 'applied physics', as Michael Heidelberger puts it. In this respect, Hering is taking up another idea of Fechner's, namely that living matter is characterized by self-organizational capacities. While self-organization for Fechner is a mechanical issue, since he postulates a kind of motion that is unique to organic matter, for Hering self-organization is a feature of metabolic processes in living systems. Energy is either assimilated or dissimilated in order to preserve an autonomic equilibrium in the nervous system. Hering therefore limits self-organization to living matter, treating it as a chemical problem rather than a mechanical problem arising from different kinds of motion, as in Fechner, thereby justifying the autonomy of physiology. The appeal to a *self-organizational capacity* of living matter returns in Piaget, as we will see shortly.

2.2.2.3. Piaget's Parallelism

In the 'Introduction à l'Épistémologie Génétique', Piaget advocates parallelism definitively as a heuristic for psychology, formulating it as follows:

> Every mental event has a physiological concomitant[13].

He says that the acceptance of this empirical postulate in the nineteenth century made it possible for people with different ideas about the relationship between mind and body to put philosophical issues aside temporarily and establish an empirical basis on which philosophical issues could later be taken up again, and also enabled psychologists to free psychology of metaphysics in order to make it into a respectable science. He also offers a provisional interpretation of it, which, like Fechner's interpretation, denies a causal interaction:

> There exists no connection (causality, interaction, etc.) between mental and physiological phenomena, other than precisely that of concomitance[14].

This distinction between the empirical postulate and the non-causal interpretation corresponds exactly to Fechner's distinction. But Piaget seems not to have been directly influenced by Fechner. For one thing, he never mentions Fechner, although he loves to name-drop. Secondly, he speaks of Weber's law when he means the Weber-Fechner law. He does speak

[13] Piaget 1950, 171, my translation.('Tout phénomène psychique a un concomitant physiologique déterminé.').

[14] Piaget 1950, 171, my translation ('Il n'existe aucun lien (de causalité, interaction, etc.) entre les phénomènes psychiques et les phénomènes physiologiques, sinon précisément de concomitance.').

very favorably of Hering when discussing the role of physiology within psychology, and refers to him somewhat regularly.
For example:

> And, contrary to the intellectualist tradition that runs from von Helmhotz via the Graz school and Meinong to von Weizsäcker, the tradition from Hering to modern Gestalt theory is characterized by the consistent recourse of perceptual psychology to physiology[15].

So a direct influence from Hering is likely. Piaget also cites Theodor Flournoy (1854-1920), with whom he studied philosophy in Paris and with whom he was planning on writing a dissertation. Flournoy defended an instrumentalist philosophy of science that drew on Mach. Piaget also mentions Harald Höffding (1843-1931) as an authority on parallelism. He was in all likelihood also exposed to the notion while working under Théodore Simon (1873-1961) in the lab founded by Alfred Binet (1857-1911), who had a correspondence with Mach and who discusses parallelism in a rather Machian way in a section of *L'Ame et le Corps* entitled 'Le parallélisme'. Binet mentions the threat of epiphenomenalism as a main problem with parallelism. He also says that parallelism misleadingly suggests that a conscious phenomenon is a whole object (*tout complet*) and proposes thinking of consciousness as a mode of activity:

> Reality shows that every phenomenon of consciousness consists in a mode of activity, an ensemble of faculties that need an object in order to be applied and realized, and that object is provided by matter[16].

It seems to me that this interpretation does not do justice to Fechner's formulation(s) parallelism, since Fechner is at pains to show that psychological phenomena do not constitute a unique substance. But it does suggest the direction in which Piaget is to develop an interpretation of parallelism, namely, to conceive of psychological phenomena as a kind of activity that is essentially linked to matter. This enables Piaget to think of the psychological level of description as autonomous not because of a unique perspective – at least not directly for this reason – but because it makes possible the characterization of a unique kind of connection among phenomena that are spread out over time. Specifically, Piaget will argue

[15] Piaget 1950, 139, my translation, ('Et, contrairement è la tradition intellectualiste, qui s'est poursuivie de Helmholtz à v. Weizsäcker, par l'intermédiaire de l'école de Graz et de Meinong, c'est bien ce recours constant de la psychologie des perceptions à le physiologie qui caractérise la tradition conduisant de Hering à le moderne théorie de la Forme.').

[16] Binet 1905, 231, my translation (La réalité montre que tout phénomène de conscience consists dans un mode d'activité, un ensemble de facultés qui ont besoin d'un objet pour s'y appliquer et pour se réaliser, et que cet objet est fourni par de la matière.').

that the rational coherence of behavior in light of a person's beliefs and desires can only be captured by a psychological level of description. But before I expand on this point, let me just recap the situation: with parallelism as an empirical postulate and the non-causal interpretation, Piaget is right where Fechner left off. But he wants to avoid the charge of epiphenomenalism and thus seeks an interpretation of parallelism that will ensure that there is an autonomous psychological level of description at which a unique kind of explanation can be given, which cannot be given at a physiological level. How does it work?

A mental event, such as a decision to grab an object, is not a matter of a causal intervention of the mind into the material world. Rather, insofar as it has causal consequences, this is because it has a physiological concomitant; and it is this physiological concomitant that has causal efficacy. But once the decision has been made, it has implications for my behavior that cannot be characterized in a purely physiological language. The person carrying out this action would, for example, cast aside obstacles that appeared between the object and herself, or avail herself of any tools that should happen to present themselves. These sub-actions are rational in the context of her effort to carry out the action she has begun.

A psychological level of description enables us to formulate these implications in a way that cannot be done with a purely physiological language. We can, for example, ascribe to the person a desire to grasp the object. This enables us to make predictions about the course of her action as the situation develops that appeal to her rationality, i.e. she will avail herself of tools that present themselves, she will cast aside obstacles that appear, etc. These predictions, according to Piaget, cannot be made without appealing to the person's rationality. They follow logically from her initial decision, and likewise from our psychological description of her decision. But they do not follow from a physiological description of the event in her brain that constituted the concomitant of her decision. So, for Piaget, there are two parallel series of events. At the physiological level, there is a series of events that is causally connected. At the mental level, there is a series of events that is connected by relations of rationality. Consciousness, according to Piaget, sets the values that physiological processes maintain.

> …a series of operative and pre-operative relations between concepts and values: desire, decision and realization form from this perspective two values: one characterizes the affect to be attained (desire), the other the current effect (realization), while they can be converted one into the other by a factor (decision) that can be operative, if the will plays a role, or simply regulative. Neither the will nor this regulation are causes, however, since they are limited to determining these values by means of implication[17].

[17] Piaget 1950, 175, my translation ('La série des états de conscience consiste alors, non pas en une suite de causes et

In short, parallelism enables Piaget to assert an explanatory autonomy for psychology: He writes:

> On the whole, the principal of psycho-physical parallelism assumes a scope that goes well beyond that of a simple heuristic principal. Its true meaning consists not only in establishing a concomitance between consciousness and certain physiological mechanisms, but also in reducing the former to a system of implications and the latter to a system of causes…[18]

And:

> …in the final analysis, the principle of parallelism forms, in effect, an instrument for collaboration between the two methods of thought or two languages: the idealistic language of reduction of reality to conscious judgments and values, and the realistic language of explaining mind through physiology[19].

Psychology can and must employ a language that captures these implicative relations in order to explain behavior insofar as it is rational. So, looking back at the objection to parallelism that Piaget is addressing, namely the threat of epiphenomenalism, Piaget's proposal renders the autonomous, psychological level of description necessary for explaining behavior.

It seems also to meet the other aspect of the objection as well, i.e. the charge that parallelism invokes a purely subjective perspective that does not belong in science. The way in which this psychological level is used in Piaget's proposal has nothing to do with introspection and or, at least directly, with the first-person perspective. For the rational relations that Piaget refers to are used to predict others' behavior, not one's own. They are not subjective facts but, rather, a sort of ideal model that applies to all rational actors. As it turns out, then, Piaget avoids the charge of introducing subjective elements into science, but at the

d'effets, mais en une suite de rapports opératoires ou préopératoires entre les notions et entre des valeurs : désir, décision et réalisation constituent, de ce point de vue, deux valeurs, l'une caractérisant l'effet à obtenir (désir) l'autre actuelle (réalisation), transformées l'une dans l'autre par un facteur (décision) soit opératoire, au cas où la volonté, ni cette régulation ne constituent en elles mêmes des causes, puis qu'elles se bornent à déterminer par implication les valeurs en fonction les unes des autres.')

[18] Piaget 1950, 177, my translation ('Au total, le principe de parallélisme psycho-physiologique paraît ainsi acquérir une portée qui dépasse de beaucoup celle d'un simple principe heuristique. Sa signification réelle ne consiste pas seulement à affirmer la concomitance entre la vie de la conscience et certains mécanismes physiologiques, mais encore, en réduisant la première à un système d'implications et les seconds à des systèmes de causes…').

[19] Piaget 1950, 177-8, my translation, ('…en dernière analyse, le principe de parallélisme constitue, en effet, un instrument de collaboration entre deux méthodes de pensée, ou deux langages à traduire l'un dans l'autre : le langage idéaliste de la réduction du réel aux jugements et aux valeurs de la conscience, et la langage réaliste de l'éxplication de l'esprit par la physiologie.')

price of sacrificing the inner perspective that was at the core of parallelism.

Looking at the relation between psychology and physiology in Piaget, we may at first wonder whether, as far as Piaget is concerned, psychology might be autonomous with respect to physiology, but might amount to nothing more than a theory of rationality. That would be strange, because there would be no room left for psychology to explain idiosyncratic behavioral phenomena, i.e. for phenomena that are currently explained by psychology but are not rational. But, according to Piaget, human behavior only approximately complies with rationality. Hence, it results from a mixture of physiology and rationality; and psychology has the job not only of applying a theory of rationality, but of showing how the human physiological makeup issues in approximately rational behavior. So, psychology is well advised to refer to physiology as well. In order to explain how this works I will have to say something about Piaget's understanding of development.

According to Piaget, sensory-motor schemata are the basis of cognition. The physiological processes by which stimulus information is transformed into sensations and into actions arise during childhood as means of preserving the equilibrium of the organism. Eventually, motor schemata are re-described in more abstract terms as operational schemas, which are more successful at preserving equilibrium, since they incorporate some useful principles, such as object permanence and the distinction between inanimate things and animate agents. These schemata are in turn taken up in abstract-formal schemas. The more abstract level of description ignores many local and specific features of concrete motor schemata in order to classify them according to their commonalities. These schemata incorporate logical and mathematical principles. They are therefore ideal, or normative. This sort of adjustment, or optimization, of schemata to suit the surroundings is what Piaget refers to as 'accommodation'. It is supplemented by a complementary process, namely 'assimilation', by which new objects are associated with existing schemata.

The result of this cybernetic account of cognitive development is that, although we need an autonomous psychological level of description to capture implicative relations, psychology is also well advised to pay close attention to physiology, since the implicative relations at the *psychologically* characterizable level of the formal-abstract schemas are parallel to or isomorphic with the causal relations at the *physiologically* characterizeable level of the motor-schemas. Higher-order thought processes are shaped by their physiological origins. This means that the rational structure of cognition – i.e. the implicative relations between concepts – is not captured in a psychologically purified formal language, but is shaped also by the specific physiological structures and processes that are its origin. So he says:

> …a psychological theory of intelligence is impossible without neurology, since intelligence is a systematization of processes the roots of which are in perception and motor function[20].

I would like to argue now that this is a continuation of Hering's program. That may seem strange, since Hering argues for the non-reducibility of physiology and highlights the importance of physiology in explaining psychological phenomena. Piaget, on the other hand, is limiting the role of physiology here and invoking a separate mental level of description. But this level of description is for Piaget the result of physiological processes that are, as Hering would have had it, irreducible to physics, and for the same reason, namely the self-organizational capacities of living matter. It is probably no coincidence that Piaget, like Hering, postulates two complementary mechanisms (assimilation and accommodation) that ensure the preservation of equilibrium, and that one of the mechanisms (assimilation) even bears exactly the same name.

The weight that Piaget places upon this physiological dimension of development should make it sufficiently clear that there is something wrong with the prevalent view of Piaget's account of cognitive development as intellectualist, solipsistic or disembodied. Shaun Gallagher, for example, in a characteristic expression of this view, ascribes to Piaget a 'theory of perception in which access to a meaningful external world is not direct but mediated in a process that necessitates an acquired capacity to synthesize sensations belonging to different sense modalities in a process of intellectual abstraction.' What Gallagher thinks is missing from this theory is that 'perception is less the result of an internal processing of information, and more the result of an interaction between the body and its environment.' But, as we have seen, Piaget is at pains to show how processes of 'intellectual abstraction' can be identified with physiological processes, and that their development can be explained in terms of cybernetics, which specifically emphasize the interaction between the body and its environment. One may or may not think Piaget succeeds at linking physiology and psychology, but it does not seem fair to accuse him of neglecting the body and its interaction with the environment.

The comparison to Fechner is also interesting with respect to self-organization. Both want to make room for consciousness and free will, and to maintain an autonomous level of

[20] Piaget 1950, 139, my translation ('…une théorie psychologique de l'intelligence elle-même ne se concoit pas sans un ensemble de tels emprunts à la neurologie, puisque L'intelligence n'est qu'une systématisation des processus dont les racines plongent dans le perception, le motricité etc.').

description for psychology while avoiding the need to postulate an immaterial mental substance. And both do so by asserting that there are regularities that cannot be explained by physiology, and which arise because of the self-organizational capacities of living matter. There is also an interesting difference. For Fechner, free will and consciousness are associated with spontaneity, which is a primordial feature of the universe, i.e. it is a relict of an earlier phase of the universe, at which matter had not yet acquired the habits that appear to natural science as laws. For Piaget, on the other hand, consciousness and free will are the result of the universe's evolution toward greater stability. They emerge during cognitive development, as children discover more abstract schemas that enable them to maintain their equilibrium more effectively.

Another point at which it is illuminating to compare Piaget with Fechner is the inner perspective that is at the center of Fechner's parallelism. Fechner has the advantage that he accounts for a subjective perspective in an elegant naturalistic way, but the problem is that it is unclear whether this perspective contributes anything to science or is merely epiphenomenal. As for Piaget, his corresponding irreducible mental perspective avoids the charge of epiphenomenalism, since it is essential to psychological explanation, but it is unclear whether it has much to do with subjectivity at all. We do not use it to introspect mental states, but only to grasp connections among mental states, perceptions and actions. And I can employ this perspective upon you as well as you could upon yourself.

But, on the other hand, these rational connections are only understandable from the perspective of an agent who has mastered the actions from which they are abstracted. So, in order to understand them and to employ them in giving psychological explanations, one needs to have mastered and abstracted from the same actions. And presumably, there will be some differences between individuals and cultures in the actions they have successfully employed and built upon. So there is a certain irreducible subjectivity here. Moreover, insofar as psychology can characterize the relevant actions and capture the rational relations based upon them within psychology, it will not have eliminated but incorporated a subjective perspective. Indeed, this is what Piaget thinks, since for him science is a continuation of the same developmental processes at work in the individual, and builds upon them just as they build upon each other. That is what enables him to use parallelism as his master argument in epistemology.

Piaget likes to use the notion of a circle being closed to illustrate how he wants to use parallelism within his genetic epistemology. The central problem he poses is that, in the empirical sciences, including psychology and physiology, we have to start out by presupposing norms of thought (logical principles and mathematical truths) that he says cannot be justified unless we are willing to plunge into metaphysics, which he wants to avoid. The principle of parallelism enables him to give an empirical-psychological explanation of why the application of our norms of thought in empirical science should be successful. He argues that the norms of thought (math and logic) are isomorphic to schemata that arise through successful interaction with the world during development (i.e. for models of the natural world). As he puts it:

> …the operations of thought express reality, since their psychological roots reach down into physical chemistry[21].

There are two steps here. One idea is that the most abstract norms of thought – logic and mathematics – that arise during development are shaped by their origins in the more concrete models that we have of the natural world. So our abstract norms of thought, i.e. math and logic, should be applicable to the natural world if our more concrete models of the natural world are applicable. The latter inherit the adequacy of the former. The other point is an evolutionary explanation of why any of these cognitive processes are empirically adequate. The earliest sensory-motor schemata are themselves natural processes and are adapted to the natural environment. More complex cognitive structures simply abstract from these sensory-motor schemas, and thus preserve the adaptive function that links them to the natural world.

Hence, a combined psychological and physiological account of development is meant to show why mathematics and logic are useful in investigating the natural world in empirical sciences. So it is intended to be a virtuous, closed circle. He has physiology making use of mathematics and logic, psychology making use of physiology, and mathematics and logic making use of psychology. He also sometimes incorporates other sciences in this image, but that is not relevant to the basic point here.

Essentially, it is a straightforward evolutionary epistemology. I think that this can only get him so far. What each science gains from the other in this picture is radically different. Empirical sciences gain the ability to measure and formulate precise laws with the help of mathematics; psychology could well gain insight from physiology into the inferential structure of thought; but mathematics and logic gain only an account of their psychological genesis, which is not very interesting to most mathematicians and logicians.

Philosophically, Piaget can surely get an evolutionary epistemology that pragmatically

[21] Piaget 1950, 181, my translation ('…les operations de la pensée sont capables d'exprimer le réel, en tant que plongeant leurs racines physiologiques jusque dans la matière physico-chimique…')

justifies knowledge insofar as it is useful for survival, but it is far from clear that concepts like truth and justification can be exhaustively accounted for in terms of their contribution to survival. Piaget's account contributes nothing to these central problems. Moreover, the realist overtures he makes about operations of thought expressing and corresponding to reality is unfounded, since the self-organization of organic matter he describes does not necessarily lead to representations at all, let alone to ones that correspond to the world.

2.2.2.4. Piaget and analytic philosophy

At any rate, if we take Piaget's position with respect to the mind-problem seriously, then the question arises whether he had any influence upon or interaction with contemporary analytic philosophers. The most convenient starting point for addressing this question is in fact his naturalized epistemology, since, as it happens, he and Quine had an ongoing interaction and consistently made reference to one another in highly favorable terms over the course of several decades. Quine, one of most prominent philosophical advocates of naturalized epistemology, first heard Piaget talk in 1947, and is included among the members of the editorial committee of the *Etudes d'épistémologie génétique*, published in 1957. (Piaget[1957]). He also visited Piaget in Geneva in 1960, ten years after the publication of the *Introduction à l'Épistémologie Génétique*, and gave a brief talk in which he praised Piaget highly. In the text of this talk, he explains that Piaget:

> …is motivated by the distinctly philosophical purpose of tracing out the structure and mechanism of our thought processes and of our conception of the world. There is in all this the same motivating interest that motivates epistemology; but in addition there is the rich empirical source of understanding that comes of experimenting on the developing individual and exposing actual *stages* in the development of the thought processes and concepts in question. It is a source of which philosophers have deprived themselves who, like Husserl and others, have abjured what they call psychologism. I, on the contrary, embrace psychologism most cordially, and feel that Professor Piaget's program of genetic epistemology, as he calls it, can be an important avenue to philosophical illumination (D. Follesdal and D. Quine, 2008, 271-2).

Quine's naturalized epistemology is, of course, closely bound up with the problem of radical translation, which brings us directly back into the thick of the mind-body problem. In the radical translation thought-experiment, we see how the content of the mental states and processes that we are entitled to ascribe to other people - and if we look one step ahead to Davidson, the same goes for self-ascription - is fixed by their behavior. This means that Quine provides an epistemological gloss to Piaget's model. For Piaget, complex thoughts and cognitive processes have their *developmental* basis in physiological processes and in behavior, from which they arise via abstraction. For Quine, behavior becomes the basis in an *epistemological* sense.

This Quinean connection is also interesting in light of the fact that Donald Davidson, who of course used Quine's radical translation problem as a starting point for his work on triangulation, wound up with a mind-body theory that may be considered parallelist. As Michael Heidelberger has pointed out, Davidson's anomolous monism is akin to parallelism in that both deny an ontologically distinction between mental and neural states, asserting a token-identity on the ontological level, but consider the mental level to be nevertheless irreducible. There is a slight difference: for Davidson, mental states are functionally defined and, as such, irreducible because they play a role in functional explanations. This is not the same as saying that they are irreducible because they are given to an inner perspective. In fact, it is not even clear that the two positions are compatible. In order to make them compatible, we would have to find a way to make functional states accessible by something along the lines of Fechner's inner perspective. If we could achieve that, it would be a great help to Fechner, since it would give him an explanatory role for mental states that would avert the charge of epiphenomenalism. In fact, it would help Davidson and functionalists as well insofar as they have the problem of accounting for qualia and phenomenal consciousness and that sort of thing.

It may be that Piaget's parallelism indeed suggests a way of pulling off this combination. His rational explanations are basically the same thing as functional explanations for Davidson - who is also at pains to demonstrate that mental states are linked to each other and to behavior by rationality. That we are able to give rational explanations of other people’s behavior by appealing to our own dispositions and abilities gives us a way of producing predictions that are equivalent to functional explanations – but without employing explicitly mental concepts that would be defined in functionalist terms.

In summary, the parallelist tradition inherent in Piaget’s position suggests an interesting way of making use of the idea, central to ST, that our own first-person constitution provides us with a way of understanding others that obviates the need for extra, specifically folk psychological concepts and inferences. Like the functionalism inherent in TT, it regards mental concepts as ineliminable because of the role they play in explanations. Also like functionalism, is regards them as irreducible, but here there is a different reason: it is not so much that they pick out different classes of states (i.e. by functional criteria, as in functionalism), as that their application requires a first-person perspective. Spelling out how this first-person perspective could be compatible with a naturalistic worldview has been an ongoing challenge for thinkers in the parallelist tradition. Doing so requires finding some form of access to one's own mental states and processes – or at least one's dispositions to decide and to act. As we shall see, this is a central challenge for ST as well.

But before we take up this issue any further, let's take a step back and have a closer look at the two leading theories of that attempt to account for our everyday psychological competence. We will look first, in chapter 3, at TT, and then, in chapter 4, at ST. As we will see very quickly, both camps in fact include quite diverse theories. I will do my best to demonstrate the differences among these various "sub-theories" while at the same time showing why they are grouped together as they are. After we have examined these theories in some detail, we will turn, in chapter 5, to systematic attempts to compare, combine or replace them.

Chapter 3:
Theory Theory

3.0 Introduction

In this chapter, I would like to introduce one of the two leading theories intended to account for our folk psychological competency, namely the 'theory theory' (TT), which attempts to account for this competency by postulating that we apply a quasi-scientific psychological theory in order to make sense or others' behavior in everyday life. There are two main versions of TT, namely an empiricist and a nativist version. I will discuss and compare both here, and review some theoretical and empirical arguments for and against each.

.

3.1 Theories about theories about theories…

In the introduction to their own classic collection of papers on the ToM debate, Davies and Stone (1995) characterize the basic idea of TT as follows:

> The idea here is that having a theory of mind is having a body of information about cognition and motivation that is applicable to others, just as much as to oneself. Given such a body of generalizations, one can use premises about what another individual knows or believes, in order to reach conclusions about that individual's actions, for example (p.1).

Now, if we look at this rough-and-ready definition, the first thing we are likely to notice is that it is pretty open. What does it mean to have a body of information that is applicable to oneself or to others? One could have the information implicitly or explicitly, or in other words, procedurally or propositionally. In fact, if one uses simulations to predict others' behavior and has no explicit or propostional knowledge of how the simulation is working, we could still say that one has information about cognition that is applicable to oneself and others – it is just that the information would be implicitly embodied in procedures that one does not have propositional knowledge of. The next sentence is therefore quite helpful, in which the authors speak generalizations and logical inferences ("premises"), because this is the minimum commitment that I want to ascribe to TT. Otherwise there is no reason to use the term theory at all, and the distinction between TT and ST collapses. So the core of TT will be the claim that folk psychology depends upon some other kind of knowledge than what we use when we ourselves are deciding, planning and/or executing our own actions. To whit, it has to be psychological knowledge, i.e. it has to be knowledge that is about the procedures that we or others undergo when deciding, planning or acting, and not just those procedures themselves.

But there is an important caveat: this psychological knowledge need not be consciously accessible, and our use of it need not be subject to deliberate control. Obviously, proponents of TT, by denying that the theory in question must be consciously accessible, thereby take on the burden of justifying their use of the term “theory”, since theories are usually the product of conscious thought and are applied deliberately. Theory theorists divide into two primary camps about how to address this issue: the first I refer to as empiricists (although their version is also sometimes called the “child-scientist theory”), the second as nativists (although their version is also sometimes called “modularism”). Essentially, the difference is that the empiricists think that the analogy to science is helpful in characterizing the developmental processes that lead to a mature theory of mind, whereas the modularists are not as interested in this process as in the end product, and think that a body of knowledge that is structured like a theory unconsciously underlies the end product. In comparing this unconscious body of knowledge to a theory, they are inspired by Chomsy’s conception of unconscious innate knowledge of syntax. I will take each version in turn, and hopefully it will become clear what each means by “theory”.

3.2 Empiricist versions

Alision Gopnik, Henry Wellman and Josef Perner are probably the three most prominent advocates of the theory theory in its empiricist form, which is based upon the idea that children develop a theory of mind in more or less the same way in which scientists develop a theory about some domain or other. In their classic presentation of the position, Gopnik and Wellman (1995) start out with a rough-and-ready characterization of the sense of “theory” that they have in mind. Essentially, they simply point out that a theory postulates “theoretical constructs” that are not themselves observable, but which help to predict and explain observable phenomena, and that these theoretical constructs work together with “systems characterized by laws or structure”. Obviously, these two components correspond to mental states and nomological psychological generalizations. The interesting move that they make is to assert that the analogy to scientific theory would be borne out if the development of social cognition, or ToM, in children corresponds interestingly to *theory change* in science. The latter involves, on their account, a characteristic progression:

1) Initially, there is a tendency to overlook or deny counter-evidence.

2) Auxiliary hypotheses are introduced ad hoc as the counter-evidence mounts.
3) Eventually, an alternative is formulated and used in a restricted fashion to deal with counter-evidence, and is subsequently seen to apply more generally.

So how does this analogy apply to children's emerging theory of mind? I will first say how Gopnik and Wellman envision this analogy, and in the next section turn to an equally influential, but slightly different, version of the child-scientist theory espoused by Josef Perner.

3.2.1 Gopnik and Wellman ("the child-scientist theory")

Gopnik and Wellman isolate three stages of development (at 2, at 3 and at 5), and characterize the theory of mind (ToM) typical of chlidren at those ages. They think that children have some vague notion of others' internal states from the start, as evinced by such phenomena as imitaiton (Meltzoff and Moore 1977) social referencing and joint attention (Wellman 1990, Tomasello 1999), but that a fairly coherent theory coalesces by about 2, which persists in the face of counter-evidence, although it is enriched with auxiliary hypotheses by about 3, until it is replaced by a distinct theory by about 5.

ToM at 2

By the age of 2, at any rate, children understand that others' behavior is driven by desires and perception. But perceptions, for the 2 year-old, are not understood as representational states but, rather, as simple "awareness of objects" Gopnik and Wellman 1995, p. 236). This awareness, for Gopnik and Wellman, is limited to farily direct causal links between the mind and the world, characterized by such generalizations as "Given that an agent desires an object, the agent will act to obtain it" and "Given that an object is within a viewer's line of sight, the viewer will see it" (237). These simple laws allow such predictions as "If an agent desires X, and sees it exists, he will do things to get it." But they do not enable the 2 year-old to predict or understand behavior via representations of counterfactual states, and they make no room for alternative perspectives or for mental states to *mis*represent. As Chandler (1988) puts it, children think of objects as "bullets that leave an indelible trace upon any mind in their path".

Some linguistic observations support this construal. Children at 2 do not use cognitive mental verbs, such as "think", "know", "remember", "make-believe" and "dream", but begin to do so around 3, and do so fluently by 5. By 2, however, they refer to desires (usually with the verb "want" in English) prodigiously. Not only are connative terms more frequent, but the frequency with which they are used to give psychological explantions of people's behavior is quite high (around 50% at 2 years of age), whereas the cognitive (belief) terms used are much more frequently used "conversationally" than in giving psychological explanations (Cf. Wellman 1991). That is to say, they say a lot of things like "I know" and "you know what?", that do not necessarily reveal an understanding of the role of belief states in causing people's behavior. Accordingly, 2 year-olds give explanations of others' behavior in terms of desire in cases where older children (and adults) would refer to beliefs – e.g. in reponse to the question "Why is Jane looking for her kitty under the piano?", 2 year-olds say things like "Because she wants to find her", whereas older children say things like "because she thinks it is there" (Bartsch and Wellman 1989).

Given the analogy to scientific theory, it is natural to suppose that children would make different predictions on the basis of their differing theories. This supposition is borne out by some data, showing, for example, that 2 year-olds will correctly predict others' happiness or sadness on the basis of a desire that is either fulfilled or not fulfilled (Wellman and Woolley 1990). They can also correctly predict what others will perceive in a variety of circumstances, even when the other's perspective differs from their own, as well as the actions that others will perform on the basis of what they have perceived (Wellman 1993). They fail, however, to make predictions that correctly reflect the fact that others may perceive the same object as they do but from a different perspective, and thus gain different, superficially contradictory, information about the object (Flavell, Everett, Croft, and Flavell 1981). In other words, they think of others' mental states as being transparent, unmediated by representations, rather than opaque, as we adults conceive them.

Wellman sometimes refers to 2 year-olds psychological theory as being a *desire psychology* rather than a *belief-desire psychology* (Wellman 1991). Desires are arguably unobservable mental statesbut the idea is that children's understanding of others' desires is not based upon ascribing internal representations of desired object (say, an apple) to others. Instead, they themselves represent the outer object (the apple), and ascribe to the other person a longing that is directed to the apple, and not mediated by an internal representation of the apple. This enables them to predict some aspects of the other's behavior (looking for the apple, being disappointed at not finding it, etc.) At any rate, we will see further below (in section 2.2) that the theme of regarding 2 year-olds as deriving psychological predictions from an understanding of situations rather than from an understanding of other minds recurs also in Perner's theory, and is common to theories that focus on domain-general developments as the main cause of children's improvement at false belief tasks between 2 and 5.

ToM at 3

3 year-olds understand mental representations to some extent, and even have a rudimentary concept of belief. But their new appreciation of mental representations is used only as auxiliary hypotheses are used, in isolated cases and without any overall revision of the underlying theory. The concept of belief at this stage is a mere extension of their earlier understanding of perceptions, insofar as people's beliefs are taken to track states of affairs in the world. It allows for action on the basis of presently unperceivable facts, and also makes room for others' ignorance of facts. But children at this age are not able to predict behavior on the basis of false beliefs, and expect the actor's desire to combine simply with the facts, rather than with the actor's beliefs, in order to determine action. Aside from the fact that they begin around this age to use cognitive mental verbs and to give explanations in terms of beliefs as well as desires (see above), they are also able by about 3 to account for people's (true) beliefs in predicting their behavior. So, for example, if they are told that Sam wants to find his puppy, which is either in the garage or under the porch, and also that "Sam thinks his puppy is under the porch", they will correctly predict that Sam will look under the porch first (Wellman and Bartsch 1988). Just to mention in passing: in this particular study, the same task was also presented to children in the opposite form, with a negative belief content (i.e. "Sam thinks his puppy is not under the porch"), and children were about equally succesful. This rules out the possibility that the children were employing superficial tricks, such as repeating the last location heard (echolalia).

There is also some support for the idea that children at this stage resist counter-evidence, as scientists working withtin a particular paradigm do. For example, Bartsch and Wellman (1989) showed that 3 year-olds will persist in failing to ascribe a false belief to someone even in the face of counter-evidence. Interestingly, though, if asked to explain why the other person continued to look in the wrong place, some of them invoke the person's false belief. So it seems they have the resources to give such explanations or make such predictions, but resist doing so because it does not fit well with their overall theory.

Another striking example comes from a study conducted by Flavell, Flavell, Green and Moses (1990). In this study, a 3 year-old child sees a blue cup, agrees that it is blue, and then sees it hidden behind a screen. At this point, an adult enters the room and says, "I cannot see the cup… where is the cup… hmmm, I think it is white but I cannot see it." The child is then asked what color the adult thinks the cup is, and insists that the adult thinks it is blue. This is quite shocking unless we understand that the child is still clinging to a theory in which people's beliefs track facts in the world. The belief concept has been introduced, but is not yet developed enough to accommodate false beliefs.

Aside from the classic false-belief task, there are also various permutations, such as the smarties test (Perner, Leekam and Wimmer 1987), which reveal the robustness of this specific difficulty that children face. In the smarties test, children are given a box of candy (smarties) which turns out to be full of pencils instead of candy. When asked what someone else will think is in the box (the other person has not seen the box before or looked inside), they reply that the other person will think there are pencils inside, even though they themselves have just fallen prey to the same deceptive appearance and should (one thinks) realize that the other person will make the same mistake.

It is worth noting that when children are asked what they themselves thought was in the box before they looked, they reply that they thought there were pencils in the box (Gopnik and Astington 1988), which suggests that their ToM has the same limitations when applied to themselves as when applied to others. This interpretation is supported by findings in structurally similar tests that lack the mind-to-world direction of fit characteristic of beliefs. So, for example, 3 year-olds were asked to imagine a blue dog and then to imagine a pink cat. When asked what they were asked to imagine first, they have no problem (Astington and Gopnik 1991). It seems, then, that the difficulty in the former sort of task is specific to the concept of belief. They understand that beliefs are not just any representations but representations of the world that have the function of reflecting the present state of affairs. They just do not understand that those representations can be *false*.

It is telling that 3 year-olds also fail to distinguish adequately between appearance and reality in tasks that do not directly have to do with ToM. One study, in particular, brings this out quite nicely. The children are shown a white cardboard flower, and they are shown a blue color filter, which has the effect of making whatever is seen thourgh it appear blue. The white cardboard flower is then placed behind the filter so as to appear blue. A piece of the flower is then cut off as the chidren observe through the filter. Then they are presented with a white piece of cardboard and a blue piece, both of which have the same shape, and are asked which piece came from the flower. Astonishingly, they think the blue piece must have come from the flower, even though they are no longer looking through the filter (Flavell et al. 1986).

In keeping with Gopnik and Wellman's view that desire understanding precedes belief understanding, children succeed much earlier at recalling their own unfulfilled desires than at recalling their own false beliefs. In one study, for example (Astington, Gopnik and O'Neill 1989), children are shown a box of crayons, which they like to play with, but the box turns out to contain something boring, such as candles. Most 3 year-olds will correctly report afterward that they did not get the toy they wanted, but only 54% were able to correctly report that they thought there were candles in the box, even though all of them had said they thought there were crayons in the box.

ToM at 5

By 5, children have developed a more sophisticated theory, according to which all psychological functioning is mediated by representations. Other people are seen to act not on the basis of facts about the world but on the basis of their representations of the world, which have the function of tracking facts in the world, but fulfil this function only more or less well. In other words the potential for misrepresentation has ben incorporated. The upshot is that they have developed a representational theory of mind; the Copernican revolution is complete.

3.2.2 Perner and the representational theory of mind (RTM)

Josef Perner, who pioneered the false belief test paradigm as a crucial measure of mature ToM, has been not only a leading designer of experiments but also one of the most prominent figures in the theoretical discourse devoted to interpreting experiments and constructing a coherent theoretical account of the development of ToM. From the perspective of philosophers interested in ToM research, Perner is especially relevant insofar as he has consistently engaged with philosophical theories and concepts, from Frege's sense-reference distinction to more recent work on the concept of representation (e.g. Dretske 1981, Dennett 1987, Millikan 1984, Fodor 1987). Perner starts out with a distinction among three levels of complexity within the concept of representation: *primary*, *secondary* and *meta-representation*.

At the *first level of complexity*, only a minimal criterion is fulfilled: namely that there has to be a causal relationship between the representation and that which is represented. For instance, a photograph of a horse represents a horse in part because, "as a result of the photographic process, the fact that the object photographed is a horse caused the image on the picture to take on the shape of a horse" (1991, 7). This causal connection underlies what Perner (following Leslie 1987) calls "primary representations", which are characteristic of children's relation to the world in their first year. They simply maintain a single model of the world, which only allows for the mind to register what is currently presented to it in the actual world, i.e. it leaves no room for past, future, pretend or counterfactual reality. Perner also sometimes refers to this kind of representation as a "single updating model." He thinks that children in their first year already grasp other minds as representing the world in the sense of this basic conception, as manifested by their awareness of others' direction of gaze and of others' emotional reactions to events in the world. Obviously, Perner is very much in line with Gopnik and Wellman in with respect to their view of children's earliest conception of others' mental lives[22].

In his seminal 1991 monograph, *Understanding the Representational Mind*, Perner addresses a potential challenge to this view of very young children's representational capacities. He notes that there is evidence that 3.5 to 4.5 month-old children continue to be aware of the existence of an object when it is obscured by a screen (Baillargeon 1987). This would appear to indicate that they are not limited to modeling only what is actually presented to them in the present situation. Perner's counterargument is that this finding is in fact compatible with the claim that these children have only one model, which is based on perception in the present situation. The trick is that the invisible part of the situation (i.e. where the occluded object is) is *not updated* when it is occluded, so it remains in the model despite the fact that it is not presently accessible to perception. Looking at it another way, there is no new perceptual information about the occluded part of the situation that would make an update necessary, so the model continues to reflect the state of affairs given by the most recent perceptual information from that part of the situation (Perner 1991, 46).

In their second year, Perner believes that children reach a *second level of complexity* in their understanding of representations. The crucial feature of the conception emerging at this stage is the possibility of simultaneously having numerous representations of the world, and thereby of engaging in pretence play and of thinking about the past and the future. Moreover, their ability to understand means-ends relationships, which require the coordination of multiple models of the world (i.e. at different stages in the process in question) improves dramatically around 1.5 years.

They can distinguish a picture of Daddy from Daddy, and know that the person represented is the same. But there is "no mental model representing the picture as a physical object representing, for instance, Daddy." (73) So the situation – the real one and the one in the picture – do not have any internal semantic relation to each other. They are simply two

[22] i.e., that others' directly register whatever information there is to be had from objects in their line of sight but cannot conceive of alternative perspectives or misrepresentation (see above, sec. 3.2.1).

separate situations that happen both to involve Daddy. They can coordinate multiple representations of different situations, although they do not yet grasp the representational relation as such. Perner therefore refers to children beginning around 2 as "situation theorists" as opposed to "representation theorists". This facility with multiple models begins to manifest itself around 2 and improves gradually until it issues in the mature understanding of representation beginning around 4.

This emerging ability to negotiate multiple models is revealed, for example, in an experiment conducted by Deloache (1987, 1989). Children were shown two rooms, one a smallish laboratory room, the other a miniature model of the first. They were told that the first was Daddy Snoopy's room and the second Baby Snoopy's. The two rooms contained all the same objects in all the same places. The children were also told that Baby Snoopy and Daddy Snoopy like to do the same things, and this was demonstrated to them with a few examples (e.g. Daddy Snoopy sits in his armchair, so does Baby Snoopy, etc.) Then they watched as Daddy Snoopy hid in the cupboard, and were told that Baby Snoopy was hiding in the same spot in his miniature room. The test question was "Where's Baby Snoopy hiding?" The result was that the older children in the study (slightly over 3) had no problem with the question, but the younger children (2.5) had no idea how to respond. Perner does not think that success at this task reveals an understanding of representation, i.e. that the miniature room would represent the larger room. Instead, the children have merely learned to grasp correspondences between the two situations and to draw inferences on the basis of them.

Around 4, and certainly in most cases by 5, children attain to the *third level of complexity* with respect to their understanding of represenations. The decisive feature of this third level of complexity, on Perner's account, is that representations are understood as representations in – in other words, they are *metarepresented*. This means that others' epistemic states are conceived of as attempting to register objective states of the world, and as running the risk of failing to do so, i.e. of misprepresenting the states of the world in question. But, let me emphasize that Perner conceives of this decisive new competency as a general skill that applies to children's own mental states, to others' mental states, and also to other non-mental representations, such as photographs as linguistic representations[23]. I mention this because it distinguishes Perner – along with Gopnik and Wellman – from the modularist

[23] If you bristle at these representations being called non-mental, you are free to think in terms of the distinction between inherent and derived intentionality instead. Thoughts are inherently intentional, representing things in the world (or in counterfactual or fantasy worlds, etc.), whereas language, arrows, pictures, derive their reference, or intentionality from our interpretations of them, and are this not inherently representations (Searle 1983).

faction within the TT camp, who think the mechanisms underlying ToM are specific social-cognition-resources as opposed to more general resources like the ability to metarepresent.

Working together with Johannes Brandl, Perner has developed a more specific functional account of how metarepresentation helps with ToM than he was able to give in in the 1990s, so I will focus on this more recent version (Perner and Brandl 2005). Since this account is inspired by Heim's (2002) file change semantics, let me introduce this idea briefly. File change semantics is simply a functional account of how the relevant referents are kept track of in the context of some discourse or other, and how new information about those referents is ascribed to the proper referents. The account is framed in terms of a filing system that each participant in the discourse would use, in which the referents are represented on file cards. I will borrow the same simple example from Heim (2002, 226) that Perner and Brandl use: I hear the two sentences (a)"A woman was bitten by a dog" and (b)"It jumped over the fence". In the discourse so far, then, there are three referents: "a dog", "a woman" and "a fence". The idea is that, for each referent, I introduce a new file card[24], and encode on it whatever information I have about the referent.

Card 1	Card 2	Card 3
-is a boy	-is a dog	- is a fence
-was bitten by 2	-bit 1	-was jumped over by 2
	-jumped over 3	

Okay, so applying this apparatus to the false belief task, which is our crucial test of metarepresentational ability in social cognition, we get the following. In a typical false-belief scenario, you have two characters, e.g. Max and Mary, and you have a chocolate bar as well as two locations, such as a drawer and a cupboard. To keep things simple, let's focus on Max and the chocolate bar. So here is what a child around 3 can come up with:

Card 1	Card 2	Card 3	Card 4
-is a boy	-is a chocolate bar	-is a cupboard	-is a drawer
-is searching for 2	-is in the 3	-contains 2	

[24] There are some fine points that I am ignoring here: strictly speaking, a new file card is introduced for each indefinite expression but not necessarily for definite descriptions; names are also a bit complicated. But all this is irrelevant to the main point.

What the 3 year-old child does not have in this system is room for alternative perspectives upon the same set of objects (i.e. the chocolate bar, the drawer, and the cupboard). Once she has encoded the location of the chocolate bar, there is no additional slot for *Max's* belief about its location. An older child, however, can manage to maintain a subset of file cards that stands for someone else's file card system. So cards 1-4, above, would all be filed under Card A, which would have the additional superordinate information:

Card A
-is a point of view
-belongs to me

Additionally, there would other such superordinate cards for other actors, such as Max:

Card B
-is a point of view
-belongs to 1

Under card B, the following cards would be filed:

Card 1*
-is a boy (**is the same as 1**)
-is searching for 2

Card 2*
-is a chocolate bar (**is the same as 2**)
-is in 4

Card 3*
-is a cupboard (**is the same as 3**)
-is empty

Card 4*
-is a drawer (**is the same as 4**)
-contains 2

Having subsets of cards representing alternative points of view enables the older children to negotiate between different representations (via the identity assertions that appear in boldface) of the same referents, and thus to incorporate conflicting information (e.g. true and false information) into their reasoning about those referents. What this functional analysis illlustrates, essentially, is that one needs to be able to represent the representations of the relevant objects and actors in order to make sense of false beliefs.[25]

By way of a bit of evidence for this analysis of the developmental change that enables chidlren to succeed at false belief tasks, Perner and Brandl show that the same appeal to file change semantics provides an explanation of why success at the so-called “alternative naming game” emerges in parallel with success at false belief tasks. In the alternative naming game, the child is taught two words for the same object, such as “bunny” and “rabbit” or “puppy and “dog”. When the other player uses the one word, the child is supposed to use the other. Children under 4 cannot get this right, and success at the task correlates quite well with success at false belief tasks (Perner and Brandl 2005). Note that it makes no difference whether synonyms are used, or words that stand in sub-/superordinate relation to each other, such as “animal” and “dog”. It does, however, make a big difference if both words can be construed as referring *partly* to the object, i.e. in the manner of compatible properties being ascribed to the object. For example, think of a football that is partly green and partly blue. 3 year-olds can successfully name the color that another player does not name, i.e. if the other player says the football is “blue”, a 3 year-old will be fairly successful at saying it is “green” when it is her turn (Perner et al 2002). The difference between the two games, on Perner and Brandl’s analysis, is that in the the color game the 3 year-old can encode “partly blue” and “partly green” on the same file card, and therefore has no need to formulate an identity assertion that metarepresentationally targets two separate file cards. Hence, she can get along with just *one* file card:

Card 1
-is a football
-is partly blue
-is partly green

The alternative naming game, however, cannot be mastered without the ability to formulate identity assertions such as the one depicted here:

Card 1 (is the same as Card 2)	Card 2 (is the same as Card 1)
-is a dog	-is an animal

[25] File cards occurring within different subsets (i.e. points of view) which stand for the same referents and which are accordingly identified via the identity assertions can be thought of as Fregean senses, if one wants a comparison to familiar philosophical terminology.

-is in location L	-is in location L

And formulating this identiy assertion presupposes the same metarepresentational ability as in the case of the false belief task (Perner and Brandl 2005).

Insofar as the parallel development of competency at these two games calls out for explanation, the explanation in terms of metarepresentational ability, spelled out in terms of file change semantics, can be said to gain in plausibility by virtue of the fact that it applies equally well to both cases. And since the linguistic task (alternative naming) is at least not directly a case of social cognition (perhaps this could be debated), the account given here speaks in favor of Perner, Gopnik, Wellman as opposed to Leslie, Carruthers and other modularists. Needless to say, the file card system is a mere analogy that helps in giving a functional description of the mechanisms at work. But one would like to hear something about how and where such a system is likely to be realized in the brain, especially since there are likely to be numerous functional accounts that are empirically more or less adequate, so compatibility with neuroscience would be a good criterion for deciding among them. That said, there have not been that many alternative functional explanations that have been as well formulated and empirically motivated as Perner's, so for the time being his account certainly deserves to be considered one of the leading options.

Perner's account is closely related to Gopnik's and to Wellman's. They all argue for an initial understanding of others' mental lives that is based upon a simple understanding of representation, which improves over time until a fairly mature ToM is in place by about 5. Aside from the fact that Perner focuses more closely on the concept of representation (and metarepresentation), there is also another crucial difference between Perner and the other two. Specifically, Perner does not accept the theory-shift account of development put forward by Gopnik and Wellman. He thinks that the earlier theory, situation theory, remains the default option in most cases, and that we adults are still basically situation theorists at heart. It is just that the more sophisticated understanding of representation that is in place by about 5 enables older children and adults to think in terms of metarepresentation in the special cases where is it appropriate. Perner therefore thinks that theory extension is a more useful metaphor than theory change, and points to the enrichment of classical genetics by the discovery of the structure of DNA as a better example than Copernicus-Kepler, since classical genetics is still quite useful for many purposes, just as situation theory is satisfactory for many everyday cases of understanding others' intentional action.

3.2.3 Criticism of empiricist versions

From within the ranks of the theory theorists, Leslie and German (1995) have levelled fairly harsh criticism of accounts on which the emergence of the ability to metarepresent is responsible for children's improvement at false belief tasks between 2 and 5. This is targeted primarily at Perner, but also applies to other domain-general accounts of the changes leading up to success at false belief tasks – i.e. such as the accounts put forth by Gopnik and Wellman. Some of this criticism is theoretical and some empirical.

Theoretically, Leslie and German (1995) assert that the empirical evidence marshalled in support of RTM does not in fact support it over versions of TT that treat beliefs as propositional attitudes rather than as representations. Specifically, they deny that children's success at false belief tasks by about 4 reveals the emergence of new metarepresentational skills. The falseness of a belief, they argue, is a semantic property, and is as such irrelvant to distinguishing beliefs qua representations from beleifs que propositional attitudes. To see the difference between these two options, consider the fact that one and the same belief could be represented in various ways – e.g. the belief that the cat is on the mat can be represented by a sentence or by a visual image. In both cases, the belief has the same semantic content even though it is realized by different representations. The criticism, then, is that there is no reason to think that the emergence of the ability to ascribe false beliefs has anything specifically to do with representations.

As Leslie and German construe Perner's position, to claim that 4 year-olds begin to regard others' beliefs as representations is tantamount to asserting that they begin to individuate others' beliefs on the basis of their formal, or syntactic, properties rather than on the basis of their content. Perner's position would only be supported, then, by evidence showing that 4 year-olds distinguish between mental states that have the same content but are represented differently. Leslie and German note that no such evidence has been produced, so there is no reason to think representational skills are the central issue

The empirically motivated criticism turns upon so-called false photograph tasks. In the basic version of false photograph tasks (e.g. Zaitschik 1990), children watch as a photograph is taken of a scene, then observe as the scene is changed. Meanwhile the photograph is being produced. Then the chidlren are asked to describe the scene that will be depicted in the photograph once it is ready. The crucial question is whether they can deal with the fact that the photograph will depict the scence as it was at the time when it was taken rather than as it is now. This task has the same structure as the false belief task, but does not target *mental* representations.

According to Leslie and German, Perner's position predicts that this task should be slightly easier than understanding false beliefs, since false beliefs tasks involve all the difficulties of the false photograph task plus some additional difficulties. I must say that I do not find this all that clear-cut[26]. Accounts focussing on general cognitive abilites, such as the child-scientist acccount and the RTM, should tend to expect a more or less simultaneous development of children's abilities to pass these two tests. Explaining the ability to ascribe false beliefs by appealing to general metarepresentative skills just amounts to saying that there are no "additional difficulties" in the case of mental states. Of course, advocates of these accounts would have the option of postulating "additional difficulties", but it would be an ad hoc move. That said, it is at least clear that Perner and co. have no reason to predict that children would be succssful at false belief tasks *before* they are successful at false photograph tasks. And, as Leslie and German, point out, there is some evidence that this is the case. But I am getting ahead of myself; let me reveal the results.

As it happens the children in Zaitschick's (1990) study fared about equally well with false photos as they did with false belief tasks, with 3 year-olds having lots of trouble and two-thirds of 4 year-olds getting the test question right. In a follow-up study conducted by Leslie and Thaiss 1992), it was found that, despite the prevalent correlation of success at false belief tasks and success at false photo tasks, if a child passes just one and not the other, it is the false belief tasks that they pass first. This result does tend to suggest a domain-specific explanation for the proficiency with false beliefs. Moreover, autistic children do quite well at false photograph tasks although they struggle with false beliefs. This, too, supports a domain-specific explanation of children's development of a ToM. There has been plenty of work on the experimental paradigm over the past 10 years, and we will come back to it once we have introduced the nativist approach.

3.3 Nativist versions

Another popular subset of theory-theories emphasizes the tacit nature of the theorizing involved in folk psychology, and tends to lean upon the analogy to Chomskyan linguistics in order to clarify what role could be played by tacit theory in cognition. Stich and Nichols (1995) characterize this sort of Chomsky-inspired appeal to tacit theory as the "dominant experimental strategy" in cognitive science. This strategy, they explain,

[26] Perner is not explicit about this, but he seems to agree with me. I will discuss what he says in the general discussion at the end of the chapter.

> …proceeds by positing an internally represented 'knowledge structure' – typically a body of rules or principles or propositions – which serves to guide the execution of the capacity to be explained. These rules or principles or propositions are often described as the agent's 'theory' of the domain in question. In some cases, the theory may be partly accessible to consciousness; the agent can tell us some of the rules or principles he is using. More often, however, the agent has no conscious access to the knowledge guiding his behavior (Stich and Nichols 1995, pp. 35-36).

As one would expect, it is natural for nativists to appeal to Jerry Fodor's conception of modularity. On Fodor's account (1983), many low-level cognitive processes, such as visual perception, qualify as modular in the sense that they fulfill certain criteria (at least to a significant extent). The most important criteria are:

1) domain specificity: there is a set of algorithms that operate only on input within a particular domain and not on input in other domains (e.g. algorithms that operate on visual input alone, or on multimodal input about a specific class of objects, such as conspecifics' expressions of emotion).

2) informational encapsulation: the operation of these algorithms is not influenced by information acquired by other means (a classic example of this is that you will flinch if I move my finger toward your eye even if you know I will not pluck your eye out since this knowledge does not reach the informationally encapsulated reflex mechanisms).

3) mandatory, or automatic, operation: one does not have to, in fact one cannot, control modules consciously (cf. the example of your flinching automatically when my finger approaches your eye.

Some other criteria that are a bit less central but also typical of modular processes: they operate quickly, since they need to consult only a restricted database; their outputs are "shallow", or simple, since the information they can encode is limited according to the information they take in as input and the algorithms that operate upon that input; their workings are mostly inaccessible to consciousness, like the deep structures of Chomsky's postulated innate grammar; they have a typical, largely invariant ontogeny, or developmental pattern; they have a fixed neural architecture.

Insofar as nativists think of the acquisition process as a kind of maturation or triggering, one might say that they appeal to science only to articulate the end product of the development of ToM (i.e. to the theory), and not to articulate the development of the theory. In other words, there is no *theorizing* that gives rise to the theory, as there is in Gopnik and Wellman's account. This is in fact the crucial difference between empiricist and the nativist versions of TT. As Segal (1996, p. 152) observes, the former make no special distinction between the acquisition of ToM and the acquisition of other theoretical knowledge (e.g. folk physics, folk biology). Nativist accounts, on the other hand, think that there is a domain-specific set of principles that govern the development of ToM.

3.3.1 Carruthers

Carruthers, one of the philosophers who has been most active in this debate, gives a couple of theoretical arguments for nativism that are based more or less on Fodor:

1)If children acted like little scientists, gathering data and constructing a theory empirically, then it is a wonder that they should all wind up with more or less the same theory. Hence, their theory is innate: it must develop over the course of a few years, of course, but the way in which this happens is more like a maturation or triggering process than a learning process. Note that this argument depends upon intercultural invariance of folk psychology, which Carruthers asserts (1996, p.23), but the data on this is not decisive[27]. The argument could also be met by acknowledging that there is an innate "starting-state" theory, which then undergoes changes over the course of cognitive development. One may be inclined to think this is a cheap response that covertly avails itself of exactly the sort of innate knowledge that modularists claim, but this would not be a fair interpretation. For proponents of the empiricist version of TT can go on to say that the way in which the theory develops is akin to the way in which scientific theories develop, rather than being a mere matter of triggering or maturation. Indeed, Gopnik and Meltzoff repeatedly say things just like this (Gopnik 1996, p.171, Gopnik and Meltzoff 1993, p. 236).

2) The other argument is that they could not get started at all without already possessing a ToM, since the very processes of data-gathering (i.e. describing actions) and forming hypotheses presuppose a ToM. This argument for nativism is an adaptation of Fodor's famous argument for nativism about concepts, which will be discussed in chapter 7.

[27] See below , section 3.4.

Note that Carruthers allows for some simulationist components in combination with theory theory. Specifically, he thinks that the theoretical apparatus he postulates contains general theoretical knowledge that is not content-specific. This makes good sense insofar as the ToM he postulates is innate, and content-specific mental states tend not to be great candidates for innate knowledge – e.g. the belief that *The Third Man* is playing at the *Burgkino* on Tuesday. So, Carruthers admits that simulation may play a role in deriving fine-grained predictions about thoughts, feelings and actions of other people. In this respect, interestingly, he is not too terribly far away from Jane Heal's version of ST. He just refuses to accept that simulations could work without the background theory, and especially resists the notion that we could have self-knowledge without such a background theory.

One obvious challenge that arises for Carruthers, as well as for any theory theorist, is that there does seem to be an assymetry between ascription to others and ascription to oneself, i.e. privileged access to one's own mind. How can a theory theorist avoid the implausible view that one gains knowledge of one's own mental states by observing one's behavior, forming hypotheses, making predictions and chekcing to see whether they work out?[28] Carruthers' answer is twofold: for non-occurrent states, he simply bites the bullet and supposes that we do go about self-asciption in just such a way. In later writings[29], he improves (in my view) upon this point by noting that in our own cases we have more data to go on that in the case of other people. Similarly, we understand and predict our friends' behavior better than that of strangers since we have more data on them. This is quite reasonable, but I still find it a bit odd to ignore that we have episodic memory in the first-person case but not in the second- or third-person case. As for occurent states, he suggests that there could be an unconscious mechanism that gives rise, in the case that we have mental state M, to the belief that we have mental state M. This is a bit vague, and in certainly cannot be generally true, since it would initiate an infinite regress (every belief about my mental states would give rise to the belief that I have that belief and so on…). The proposal seems to imply that a whole lot of beliefs get created that do not seem to serve any purpose. If I have the belief that George Washington was the first president of the US, what is the point of also having the belief that I have this belief? At any rate, this line of thought needs work.

[28] Don't get me wrong: it is not implausible to say that we do this sort of thing sometimes; it is implausible to say that this is all we can ever do to gain self-knowledge.

[29] Cf. Carruthers 2009

3.3.2 Leslie and Baron-Cohen

Alan Leslie and Simon Baron-Cohen are probably the two most prominent psychologists advocating the nativist version of TT. Leslie (along with Baron-Cohen and others) regard autism as a source of empirical support for nativism, i.e. for the claim that ToM depends not on domain-general skills such as the capacity for metarepresentation but on a specific, modular set of abiliites. The reasoning is that autism appears to be an example of dissociation of social cognition from other cognitive skills. In a study conducted by Baron-Cohen, Leslie and Uta Frith (1985), autists were compared to normal 4 year-olds and to children with Down's syndrome. All the children involved had a mental age of 4 (as assessed by a non-verbal test). While almost 90% of the children in the non-autistic group passed the false belief task, 80% of the autistic children failed it. They (Baron-Cohen et al. 1986) also conducted a follow-up study, in which they found that autists with much higher mean IQs than Down syndrome children (82 to 64) did much worse on false belief tasks.

The argument, then, is that if autists have *specific* difficulties with social cognition, then social cognition depends on specific skills. Hence social cognition cannot be explained by appealing to the development of general skills, such as the abilitiy to understand representations as representations generally. In fact, Leslie thinks that pretence play in 2 year-olds already reveals metarepresentational skills. But it is important to note that Leslie means something very different from Gopnik, Wellman or Perner when he talks about metarepresentation.

So what does he mean by metarepresentation? He develops the concept specifically with a mind to theoretically grasping the commonalities between pretence play and other ToM skills, such as belief ascription. His view (e.g. 1987, 1988, 1991) is that pretence play already draws on basically the same ToM skills as false belief tasks, even though the former emerges by the age of 2 and the latter not until at least 3.5. Obviously, that discrepancy will need some explaining, and Leslie has a response, which I will come to in just a moment. But first let me flesh out this idea of pretence play as a manifestation of ToM.

For starters, Leslie points to an interesting fact: pretence play and the ability to understand others' pretence play arise at the same time, namely around 2. Once they begin to engage in pretence play, they also become able to distinguish between adults who are privy to or involved in the pretence world and those who are not. This betrays an ability to understand others' divergent representations of the same physical objects, which does indeed sound a lot like ToM. Leslie pursues this link by emphasizing some similarites between semantic properties of reports of mental states and characteristics of pretence play. Reports of mental states are referentially opaque, their truth value is independent of the truth value of the embedded sentences, and they can be true even if they include a singular term that does not have a reference. Pretence play, in Leslie's view, invokes the same underlying skills, as evinced by such phenomena as object substitution, attribution of pretend properties and imagining objects where none really exist. For example, in pretending that a banana is a telephone, mommy substitutes a telephone for the banana in the context of the game. The sense that the physical object, or physical reference of my attention and discourse, takes on, is therefore opaque. He comes, then, to the following definition of metarepresentation: "an internal representation of an epistemic relationship (PRETEND) between a person, a real situation and an imaginary situation (represented opaquely)" (Leslie 1991, 73). So, in our banana example, mommy is the person doing the pretending, the banana corresponds to the real situation, and its being a telephone corresponds to the imaginary situation that is represented opaquely.

Notably, autists do not tend to engage in pretence play. Leslie's hypothesis is in fact that their metarepresentational skills are impaired.

> Autistic children are impaired and/or delayed in their capacity to form and/or process metarepresentations. This impairs (/delays) their capacity to acquire a theory of mind (Leslie 1991, p.73).

Okay, then, the obvious question is: why can't children succeed at false belief tasks until years later? The answer that Leslie gives to this builds upon Baron-Cohen's of the information-processing apparatus at work in ToM, so I will strat out by sketching Baron-Cohen's conception and then turn to Leslie's answer to this question. Taken together, Leslie and Baron-Cohen have offered probably the most specific account of the information-processing mechanisms involved in the development of ToM. They regard this as a central virtue of their position, pointing out that Gopnik, Wellman, Perner and co. have not been very specific in terms of the information-processing mechanisms supposedly responsible for theory development in children.

Information-processing mechanisms

Baron-Cohen proposes four distinct modules that contribute to mindreading. One of them, the eye-direction detector (EDD), is already manifest in the neonate. Another, the intentionality detector (ID) is (according to Baron-Cohen) definitely up and running by about 6 months, and maybe already in the neonate (the findings are inconclusive). A third, the shared–attention mechanism (SAM), does not manifest itself until around 9 months. Fourthly, there is the theory-of-mind mechanism (ToMM), which manifests itself around 2 years of age, but cannot be fully employed until 4 or 5, for reasons that Leslie gives and I will come to in a moment. It is worth noting that the later manifestation of the latter two or three is no problem for the modularist account, since there is no need for a modularist to commit to the proposition that innate modules are present at birth; a modularists must only commit to saying that the development has minimally to do with learning or cultural influence, and that the form it takes does not depend on the influence of other cognitive skills, resources, modules, etc. Le me say a bit about each module in turn.

As for the EDD, its function is twofold: to pick out eyes among other stimuli, and to follow the direction of gaze to a seen object. There are very robust findings suggesting that infants already take a special interest in eyes, distinguishing them from other stimuli, looking longer at them (Baron-Cohen 1995, 39). By 6 months, infants distinguish between faces directed toward them without the eyes also being directed toward them, and faces that are not only directed toward them but with the eyes also directed toward them.

The idea behind the ID is based upon experiments that show that infants appear to recognize intentionality in certain kinds of movements, as evidenced by their expectations concerning the subsequent trajectories of these movements. This sort of phenomenon has been observed at least since classic studies by Heider and Simmel (1944) and Michotte (1963), and has been studied intensively in recent decades[30]. A typical set-up in which this phenomenon can be observed: children (in this case 9 month olds) see a dot on a screen moving in an apparently goal-directed manner, detouring around an obstacle to reach a specific target. The children dishabituated if the obstacle was removed (making the detour superfluous), but remained habituated in the face of alterations in the dot's trajectory as long the dot pursued the most direct possible path to the target (Gergely et al. 1995). It must be noted that six-month olds were also tested in the same study and did not exhibit the sensitivity to intentional motion. This suggests that the module emerges only at around 6 months. But there is other evidence suggesting that it is present in neonates (Reddy 1991). But this is not the decisive point. Regardless of when it manifests itself, Baron-Cohen is happy to regard it as innate, since he thinks that it emerges according to its own developmental timetable independently of the infant's interactions with the social environment (Baron-Cohen 1995). Others, for example Tomasello (1999), dispute this interpretation, pointing out that a host of behaviors (gaze following, understanding intentional action, and joint engagement) "emerge in close

[30] For a review of recent studies, see Scholl and Termoulet 2000.

developmental synchrony and in a correlated fashion between nine and twelve months" (Tomasello 1999, 67). In this context, Tomasello also notes that children's understanding of their own agency becomes more sophisticated in the months leading up to these social-cognitive developments, and they begin to carry out more complex actions (e.g. removing obstacles, employing tools). This suggests that their refined understanding of their own agency may provide the basis upon which they begin to understand others, and therefore supports a simulationist interpretation (Tomasello 1995, 70-77).[31]

Baron-Cohen notes that these mechanisms may well be present in other primates. Both human infants and other primates experience physiological arousal when they detect another pair of eyes gazing at them. In the case of human infants, this arousal is clearly positive, and reliably triggers smiling (1991, 42). Tomasello would add at this point that human infants seek out eye contact, which also suggests a positive emotional component, whereas other primates do not. Thus, one interesting difference between humans and other primates *may be* that primates get nervous, interested or attentive, but do not take any special pleasure in eye contact, whereas humans do.

The SAD may well also be uniquely human, in Baron-Cohen's view, and he thinks also that autists may be lacking this component (1991, 44). The SAD enables the infant to build triadic representations including Self, (other) Agent and Object. Thus, she can be aware that she and some other agent are jointly looking at the same object. Obviously, this is immensely helpful for cooperation and for learning from others. The evidence is quite robust that children around 9 months monitor the gaze of adults, checking to make sure that they are looking at the same object (e.g. Baron-Cohen 1991, Butterworth 1991, Tomasello 1999). Moreover, declarative pointing emerges around the same time. Lizskowski et al have produced especially impressive evidence that young children (at the very latest by 12 months) point at objects and are only satisfied if the adult in question looks at and can in fact see the object, and, looking back at them, responds in an appropriate way (Lizskowski, U., Carpenter, M., Henning, A., Striano, T., & Tomasello, M. 2004).

The ToMM, which is to some extent in working order by about 2, is responsible for grasping others' behavior in terms of more sophisticated mental states than is possible with only the other mechanisms (which deploy representations of goals, perhaps desires, and perceptual states such as seeing). In particular, the ToMM represents epistemic mental states, such as pretending, thinking, knowing, believing, imagining, dreaming, guessing, and deceiving. Moreover, it also ties together perceptual, volitional and epistemic states to obtain

[31] More on this in Chapter 5

a "coherent understanding of how mental states and actions are related" (Baron-Cohen 1995, 51).

Okay, so if the ToMM is in place at 2, and is responsible for the emergence of pretence play, why do children need an additional 2-3 years to understand false beliefs? Leslie postulates the so-called selection processor (SP), which is not yet operational in younger children (Leslie and Roth 1993, 100). It is a general device that is used in understanding complicated representational relations. In false belief tasks, this device must be used in concert with the ToMM, as charactizered by Baron-Cohen, in order to understand how specifically mental representations are related to behavior. Younger children have the ToMM in place, and thus understand others' mental states, but get tripped up by the representational complexity of the situation. One bit of evidence, mentioned above (in section 2.1.3), is that autists succeed at false photograph tasks but not at false belief tasks, which suggests that there is some more general skill as well as the specifically social-cognitive skill at work in false belief tasks. Another piece of evidence comes from studies showing that younger children do indeed seem to have some understanding of false beliefs but cannot apply this understanding in false belief tasks. Specifically, 15 month-olds have been found to look first to the location where the agent should indeed search on the basis of the false belief (Onishi and Baillergeon 2005). This finding, astonishing and still controversial for 15 month-olds, is in fact pretty robust and has often been replicated for 3 year-olds, who look to the correct location but then still give the wrong answer.[32] This suggests an implicit understanding, which, if we are thinking in the terms of Baron-Cohen and Leslie, is perhaps present in the EDD and the ID but cannot yet be combined with the ToMM in order to make the correct prediction, since the SP is not yet in place.

3.3.3 Criticism of nativist accounts

Although the nativist account elegantly combines work on ToM with work on pretence play – not to mention autism – it also thereby takes on its central weakness, namely that is has to give an ad hoc account why 2 year-olds who engage in pretence play do not pass false belief tests. The explanaion in terms of information processing mechanisms is a noble effort to make the discussion more specific and detailed, but there is no clear empirical evidence that is is correct.

[32] Cf. Perner 1995.

Astington and Gopnik (1991) point to another issue. They argue that processing accounts, such as that espoused by Leslie, should predict proficiency in understanding one's own mind (recalling one's false beliefs and unfulfilled desires, recalling the source of a belief, distiguishing between knowing and guessing, etc.) should precede proficiency in understanding others' minds, since one should recall one's own perceptions better than those of someone else, since the the inferential structure is simpler in first-person cases for the obvious reason that there are fewer people involved, and since one does not have to suppress one's own beliefs and desires (48). This does seem reasonable, and there is little or no evidence that this prediction is correct.

3.4 General Discussion

False Photographs

As mentioned above, false photograph tasks have been cited as a prominent source of empirical support for domain-specific – and thus for nativist – accounts of the development of children's proficiency with false beliefs between 2 and 5. The argument is that the domain-general skills required for false photograph tasks and for false belief tasks are equivalent, and that the only difference has to do specifically with the concept of belief. I have already mentioned that autists fare well at false photograph tasks even though they struggle mightily with false beliefs. There is also further evidence of a dissociation between the two tasks, dervied from beavioral studies as well as from neuro-imaging. Slaughter (1998), for example, found that normal children trained at one of the two tasks do not improve at the other. Moreover, Sabbagh and Taylor (2000) found that distinct neural systems are activated by the two tasks.

But it is not a slam-dunk case that this dissociation supports a domain-specific interpretation. there could be differences in the domain-general executive skills required after all. For one thing, as Russell, Saltmarsh, and Hill (1999) have pointed out, photographs are concrete representations, whereas beliefs are abstract. This could make it easier to disengage a photographic representation from reality. In other words, one can actually see that photographic representations can differ from the way the world presently is, where one cannot directly see that other's beliefs qua mental representations of the world differ from the present state of the world. Secondly, Sabbagh, Moses and Shiverick (2006) have pointed out that beliefs have a different function from photographs. Specifically, they have the function of tracking, or keeping up-to-date on, states of affairs in the world, whereas photographs do not.

Photographs simply record a state of affairs at a certain time, and are not supposed to change along which the depicted state of affairs. As a result, they do not become *false* when the state of affairs changes.

Sabbagh, Moses, and Shiverick (2006) therefore designed an interesting series of experiments to test the correlation between executive functioning and success at false belief tasks, on the one hand, and executive functioning and success at false photograph tasks, on the other. After finding, first of all, that executive functioning does correlate with false belief proficiency but not with false photograph proficiency, they attempted to distinguish between the two possible explanations noted above (in the previous paragraph) by investigating chlidren's understanding of false, or misleading, signs. Signs have the same job description as beliefs, namely to correctly reflect present states of affairs in the world, but are concrete rather than abstract, and therefore in this respect more like photographs than beliefs. Roughly, the idea was this. Chester tells Marianne that he is going to post a sign pointing in the direction of the house to which he is going. Marianne then wanders off. Chester sets up the sign pointing in the direction of the blue house, and proceeds to enter the blue house. Then, however, he leaves the blue house and goes into the red one, but forgets to switch the sign. Marianne comes along looking for him and sees the sign. Children are asked to which house the sign points. Of course the right answer is that the sign is pointing toward the blue house, but young children had trouble with the task. The relevant point for our discussion is that success correlated with success at executive functioning tasks, which is also true for false belief tasks but not for false photograph tasks. This suggests the provisional conclusion that representations that are supposed to track the present state of the world require specific cognitive skills that differ from the skills required for equivalent tasks involving representations but without the same function of tracking the preseent state of the world.

What does all this mean for the debate between empiricists and nativists? It is a bit unclear, I would say. Insofar as this function of tracking the world is not specific to beliefs but generalizes to signs, domain-general accounts are better off, but insofar as this is still a rather specific function, and signs can be regarded as deriving their intentionality from minds, one could regard this function as domain specific.

Conflicting desires

Some support for empiricist versions comes from recent work on children's understanding of others' desires. As already noted in the discussion of Gopnik and Wellman, the prevailing view over the past 25 years or so has been that children understand desires earlier than they understand beliefs (Wellman and Woolley 1990, Bartsch and Wellman 1995, Repacholi and Gopnik 1997). If this is true (and it does seem to be), it tends to favor empiricist theories, domain gerneral accounts of the development of ToM would predict that an understanding of desires would develop sooner than an understanding of beliefs, since the representational character of desires is simpler than that of beliefs. Let me explain why that is.

Desires are at least prima facie like photographs are distinct from beliefs and from signs insofar as they do not have the function of tracking real present states of the world, but rather, of representing possible states that are presently not real but which one is motivated to make real. In other words, desires, like photos but unlike desires or signs, do not have the function of tracking present states of the world. One likes to express this by saying that beliefs have a mind-to-world direction of fit, whereas desires have a world-to-mind direction of fit (Searle 1983). If you do not see why this should make them simpler to understand, consider this: it is possible for 2 people to have different desires with respect to a state of affairs (that it come about or not) and for both to be satisfied or unsatisfied without any contradiction. I could, for example, desire broccoli while you do not desire broccoli, and we can both be satisfied. With beliefs, this does not work. The empiricists' account of children's emerging understanding of representation between 3 and 5 gives an elegant explanation of why desires, which are representationally less complex than beliefs, can be understood by 3 whereas beleifs cannot. Nativists have no grounds to make such a prediction; indeed, early understanding of desire is an embarassment to them because they then have to give an ad hoc explanation of why young children understand desires but not beliefs, if both are mental states grasped via ToM.

There has been some dissent to this prevaling view that desires are understood earlier, and it is worth mentioning (Moore et al. 1995, Lichtermann 1991, Perner et al. 2005). Theoretically, the idea is that children at 2 or 3 understand desires in terms of objective desirability, but not as truly mental states. Let me explain. In order to conceive of desires as mental states, children would, at a minimum, have to understand that others' desires may differ from their own. A study by Repacholi and Gopnik (1997) seems to suggest that children as young as 14–months understand that an adult may desire broccoli instead of cookies even though they themselves of course desire cookies rather than broccoli. But, they may construe the situation such that it is objectively desireable for adults to eat broccoli and objectively desirable for them to eat cookies. There need be no conflict, and thus no genuine subjectivity in the different desire states.

So, Moore et al. (1995) and Lichtermann (1991) set up situations in which the desires actually conflicted with each other, i.e. they were in fact incompatible with each other. For example, in the Lichtermann study, two characters are on a boat arriving at a junction in a river. The guy wants to go left and the woman right, and the boat goes to the right. The children are asked whether the woman is happy or sad and then whether the man is happy or sad. The right answer, of course, would be that the woman is happy and the man sad. The result was that children were no better at this than at false belief tasks, thus undermining the thesis of assymetry between belief understanding and desire understanding.

But the results are still quite controverisal. Rakoczy et al. (2007), for example, see a problem in the fact that the children could have thought that the man would be happy despite his unfulfilled desire to go left, since he got to travel with the woman, who was apprently his friend (and maybe he could be happy for her, or just happy to be with her). They therefore conducted a series of follow-up experiments intended to make the conflict between the desires more prominent. The idea was to introduce quarreling between the two characters in order to highlight the incompatible, mutually exclusive nature of the two desires at issue, and also to eliminate the possibility that the character with the unfulfilled desire would be happy for the one whose desire was fulfilled. Moreover, children presumably have experience of quarreling at home, so this addition could help them to draw upon experiences that would be relevant in interpreting the situation. The result was that 3 year-old children were indeed able to correctly predict that one character, namely the one with the unfulfilled desire, would be unhappy. This result strongly supports the prevalent view, according to which understanding of desires comes earlier than understanding of beliefs – indeed, this appears to hold with regard to understanding of desires as subjective states that differ from person to person and can in fact be incompatible.

So, the work on desire understanding suggests that desires are indeed easier to understand than beliefs, even when they are incompatible. And the leading explanation for this is that they are representationally less complex, or at any rate, that they do not have the same representational function as beliefs (i.e. to track present states of the world). So empiricism seems to have a slight advantage at the moment.

Cultural invariance?

As noted in passing above, in the discussion of Carruthers' argument for the nativist version of TT, it has generally been agreed that cultural invariance of the developmental pattern and timetable of ToM would speak in favor of a modular account. The basic idea is that if children act like scientists, then the theories they wind up with should exhibit a certain amount of variety – first of all, because the data they are exposed to (including specific linguistic differences) could well differ between cultures, and also because theories are underdetermined by data anyway, so there is always a variety of possible theories that are more or less equally empirically adequate. The early data certainly seemed to support cultural invariance: Gopnik and Astington (1988) showed 3 year-olds to fail false belief tasks in North America, while Perner et al (1987) produced similar results for British children, Wimmer and Perner 1983 for Austrians, and Avis and Harris (1991) for Baka children of the Cameroons.

A recent metastudy published by Liu, Wellman and Tardif (2008) has shown the issue to be a bit more complicated. The authors compared the results of false belief studies carried out over the past 20 year in the US, Canada, China and Hong Kong. The general finding was that the trajectory of development (increase of percentage of children answering correctly as a function of age increase) was very similar across cultures, but that the timetable is delayed by 2 years in Hong Kong as compared to the other three countries. This is especially interesting in light of the fact that there would have been reason to suspect that Chinese children might become proficient earlier than North American children. First of all, there is apparently a distinction made in Chinese between "believing correctly" (in the sense of thinking something to be so, ie. believing), on the one hand, and "believing mistakenly", on the other. Children are exposed to this linguistic distinction, so one might think it would help them to grasp the concept of false beliefs earlier than children learning English. Secondly, Chinese chidren are better at executive functioning tasks than North American children (yet another empirical indication that we Americans are lazy and undisciplined!), and good performance at executive functioning tasks in North America and Europe is correlated with success at false belief tasks. But also, these factors appear not to help Chinese children. And there is no explanation on offer of why children in Hong Kong are delayed. All in all, the study suggests there is a mix of culturally specific and universl factors at play, although the culturally specific factors have not yet been isolated.

There is one other point to make about this study that is relevant to the comparison between empiricist and nativist versions of TT. To whit, the authors note that nativist versions should make the following prediction: factors that are either domain-general, like superior executive functioning, or at least not specific to ToM, such as linguistic resources, should help children to develop ToM more quickly. The reason is that on nativist accounts the relevant concepts are already in place from the start, whereas their application (or children's performance with them) is held back by domain-general factors. This prediction is not borne out by the data with Chinese children. I think this interpretation is sound, and it serves to turn the tables on Carruthers.

But Carruthers' argument also makes sense, so what are we to conclude? Although the conservative step at the moment is to provisionally conclude that it is unclear which account predicts cultural invariance and which does not, I will say that Liu et al's argument is stronger than Carruthers' by virtue of the fact that it is much more specific. While Carruthers simply claims there is support for the universality claim, Liu et al point to specific data showing that a particular kind of factor does not influence the timetable of ToM development, and give a sound account of why Carruthers (and Leslie) should make a prediction that is not successful. In summary, the cross-cultural data wins a few points for empiricist versions.

Chapter 4:
Simulation Theory

4.0 Introduction

As mentioned in the introductory chapter, the basic idea of 'simulation theory' (ST) is that instead of involving the kind of psychological knowledge postulated by TT, folk psychology is a procedure by which we put ourselves into others' shoes and know somehow directly, on the basis of our own experiences, how we would act or think or feel, and then expect the same of others. In this chapter, I will discuss and compare the three main versions of ST.

4.1 Varieties of simulation

There are various versions of ST. For example, Goldman (1989) rejects only one component of TT, namely the nomological relations used to derive predictions about behavior from constellations of mental states. He does think that mental concepts need to be used in order to set up a simulation – i.e. once we ascribe a constellation of beleifs and desires to someone, we can simulate their perspective and see how we would act. Hence, the simulation obviates the need for nomological generalizations, since we embody these without having to represent them as such. But that means that he needs an account of mental concepts that does not rest on nomological generalizations. He therefore turns to introspection and hints at the idea that mental states are differentiated by their different qualitative aspects, which are accessible to introspection from the first-person perspective. Gordon's (1986) "radical simulation" theory rejects both components of TT. Gordon appeals to Evans' (1982) ascent routine:

> If you are in a position to assert p, you are in a position to assert, I believe that p. In other words, you can tack on "I believe that…" reliably without even possesing a full concept of belief.

According to Gordon, you can assert p in a simulation of someone and tack on a sort of tag to the effect that p is their assertion and not yours, and it will function reliably like an ascription of a belief. The third main proponent of ST that I will be discussing, Jane Heal (Heal 1986, 1995), is most interested in the role of simulation in understanding others' behavior as rational. She asserts that there is no theory of rationality that could replace our own intuitions about what is rational when it comes to interpreting others' behavior as rational. Her arguments are more or less aprioristic, focusing on what it means to be rational or to understand behavior as rational, which has the effect of making her position almost certainly correct but at the same time weak, since she does not engage with empirical issues. Despite these differences, there is a central insight that is common to the various simulationist accounts, namely that *we undergo some of the same processes in explaining and predicting others' behavior as we undergo when we ourselves are deciding upon, planning and executing actions, and that this overlap obviates the need for extra, specifically folk psychological knowledge.* My goal in this chapter will be to present and contrast these three basic positions.[33]

4.2 Alvin Goldman

Alvin Goldman has been one of the most prominet proponents of ST since the late eighties, and has done the most in the past ten years or so to further develop the basic idea and to engage with empirical research (such as, but not limited to, mirror neuron research). As noted above, Goldman accepts that mental concepts play a role in simulations, insofar as one must identify and categorize someone's mental states in order to set the parameters of a simulation of them, and also insofar as one must identify and categorize one's own mental state upon completing the simulation procedure before one can ascribe it to another person. But he differs from the theory theorists in that he claims that there is an introspective component to mental concepts that cannot be captured by their functional roles. Moreover, in deriving a prediction of someone's action from their input states, one employs the same procedures that the target person employs (e.g. practical reasoning), and not separate folk psychological inferential procedures. [34]

4.2.1 Critique of functionalism

Goldman's account of third-person ascription therefore depends upon an account of self-ascription. Indeed, not just because a third-person ascription involves a simulation, which in turn involves simulatively self-ascribing a constellation of beliefs and desires in order to set up the simulation, but in a more fundamental sense, namely because he is committed, as we shall see, to the view that the content of mental concepts is specified by qualia. In order to understand this point, it will be useful to look at an article in which he criticizes functionalism

[33] Paul Harris (e.g. 1995, 1996) has also been an important spokesman for ST. I think his most valuable contributions have been not so much in proposing a theoretical program as in interpreting experiments, criticizing TT and countering criticism of ST. Thus, it seems appropriate to me to discuss individual ideas of his whenever relevant, rather than dealing directly with his program as thoroughly as I intend to deal with Goldman, Gordon and Heal.

[34] Both of these points are spelled out in Goldman 2006. For the introspectionist account of mental concepts, see chapter 9.

as well as TT, specifying the latter as a “the underlying idea of functionalism”[35]. In this paper, he is primarily looking for an account of mental concepts that is psychologically realistic; i.e. putting aside ontological and epistemological issues. Goldman frames the question as pertaining to the semantic content of mental predicates. His first step is to limit the discussion to self-ascription. This is obviously a controversial move, and we will come back to it. But it is in fact essential for his argument. Indeed, he says that “ascriptions to others, in my view, are ‘parasitic’ on self-ascriptions”.[36]

Goldman assumes that a competent speaker associates a semantic representation with a psychological predicate. He calls this semantic representation a category representation (CR), since it demarcates the category that the word denotes. Obviously, there are various ways in which the CR could be instantiated in the brain (neural networks, prototypes, feature lists, etc.) Goldman wants to be neutral about this, which is fine and good. He also acknowledges that one could doubt whether this CR, which guides the speaker’s ascription, in fact constitutes the meaning of the psychological predicate in question. After all, he concedes, one could go along with Putnam and deny that speakers must have full knowledge of the meanings of the words they use (Goldman 1993, 350). But in the midst of all this fair-mindedness, Goldman brushes this objection intrepidly away with the rejoinder that the speaker must know best what guides her own use of the word, i.e. which features of the object count for her as making the object fit the CR. This is not reasonable. There is no reason why unconscious neural processes should not lead to the conclusion that an object fits a certain category. A speaker could be led to choose a word to name an object by processes she knows nothing about. She may or may not then confabulate some features that she thinks were decisive in her decision, but I see no reason why her report should be authoritative. We will see below that this sleight of hand plays a crucial role in Goldman’s argument.

There is another problem with Goldman’s setting up the discussion as he does: it is far from clear that the best way to determine whether a person employs or possesses a given concept is to test whether she employs or possesses a particular mental predicate in a natural language. By committing to explicit linguistic utterances as the sole criteria of concept employment/possession, Goldman rules out the possibility of testing by other means. This is an unjustified limitation, and it is especially infelicitous in light of the fact that we cannot make any use of his linguistic methodology in testing pre-linguistic children and non-linguistic animals like apes.

[35] Goldman 1993, p. 369
[36] ibid., 349

Okay, but getting back to his argument: Goldman targets functionalism as a psychological hypothesis about the basis upon which people self-ascribe mental predicates. According to this hypothesis, as Goldman formulates it, the CR for a mental concept must represent a set of the functional relations that the mental state has to other mental states, to perceptual input and to behavior. Self-ascription would therefore result from a match between an occurent state, or instance representation (IR), as he puts it, and the CR representing such a set of functional relations. Goldman proceeds to point out three difficulties that emerge from this general picture.

Firstly, we often ascribe mental states to ourselves in the absence of information about their causes and effects. This is fair enough. Goldman gives the example of someone who wakes up with a headache and does not know what caused it[37]. Surely this person is in a position to ascribe a headache to herself. In such cases, Goldman maintains, we must be appealing to intrinsic features of the states in question. But in the example, the categorization turns upon the sensation(s) associated with a headache. And this is not obviously the case for mental states in general. It is not so clear that I self-ascribe the mental state of believing that 7+5 is 12, for example, on the basis of intrinsic features of the state itself.

I would also note that even for the cases for which Goldman's example can stand, he does not succeed in undermining functionalism as a partial account. A functionalist could maintain that functional relations fix the meanings of mental terms, and that children learn to associate particular sensations (such as a headache) with particular words, such as "headache", partly because having the sensation in question causes them to produce behavior that others use as a basis for ascribing them a headache. Subsequently, they are able to self-ascribe on the basis of sensations without having to take into account their own behavior, but that is because they have matched the behavior to the senation and can thus use either as a basis for the ascription. The functional relations would still play a role in the psychological account of development.

The second difficulty that Goldman points out is really a follow-up upon the first. He acknowledges that a functionalist could respond to the first point by introducing subjunctive functional relations. So, if I am unaware of the causes of my present mental state, I can nevertheless consider how it disposes me to act in various counterfactual instances, and thereby home in on its functional profile. Goldman tries to turn the tables against the functionalist here. He argues that introducing subjunctive properties into the CR yields too many properties that would be among the necessary conditions for applying a concept, which

[37] Goldman 1993, 353

would make the CR too complicated and the cognitive demands too great in assessing a possible match in any given case. I am not fully convinced by this: the subjunctive properties could be dispositionally present in the CR, and worked out on the fly in a given case where an IR is being assessed. And it is surely unnecessary to suppose that all or even some wild number of the subjunctive properties of a CR must be represented. Some typical ones would suffice.

Goldman also doubts that functionalism can give an account of how a person could assess the subjunctive properties of an IR. It seems, in this case, that a simulation would be just what the doctor ordered. So it is strange that Goldman does not bring up this option. What he does say, though, is that self-ascriptions often occur so quickly that the time frame is too short for subjects to execute counterfactual imaginative routines[38]. Maybe so.

The third point is that the holistic character of functional relations leads to a regress if state A is defined in terms of state B, state B in terms of state C, etc. This is indeed a problem, and, as we will see in chapter 7, it is a problem faced by the kind of theory of concepts that functionalism and TT rests upon. I will not discuss it in detail here, but suffice it to note that it seems plausible, as Goldman argues, that at some point one has to stop defining concepts in terms of other concepts and define some of them in a more direct fashion.

4.2.2 Empirical support that Goldman marshalls for ST

As I have noted, one of Goldman's strengths has been his intense engagement with relevant empirical in neuroscience and psychology. The work I mention here can also be taken to apply to other versions of ST. To whit, all of it applies to Gordon, but none of it applies to Heal since she argues that her version of ST is true apriori, as we shall see. One can divide the empirical evidence that Goldman marshalls in favor of ST into psychological and physiological studies. The psychological studies can be divided into those which reveal an egocentric bias in prediction of others' behavior and/or decisions, and those which reveal that children's understanding of others' behavior in particular scenarios improves when they have the opportunity to experience those scenarios from the first-person perspective. Among the physiological studies, the two primary types of evidence are motor resonance phenomena (i.e.

[38] Goldman 1993, 354

mirror neurons) and paired deficits, i.e. studies which reveal a correlation between particular deficits in recognition or understanding of others' behavior or emotions on the one hand, and parallel deficits in production of the same type of behavior or experience/expression of the same emotions. First I will discuss the psychological studies.

4.2.2.1 Psychological evidence for simulation theory

The egocentric bias

One kind of study revealing an egocentric bias in folk psychology focuses on situations in which the subjects have access to more knowledge than the people they are observing. Camerer, Loewenstein and Weber (2003) report a study in which well-informed subjects were asked to predict corporate earnings forecasts made by other, less-informed people. The subjects knew what relevant information the other people were missing but nevertheless tended to incorporate that knowledge into their predicted forecasts. Newton (1990) reports an amusing study in which subjects were given a list of 25 well-known popular songs and instructed to choose one and to tap the rhythm out to a listener. When asked to assess the likelihood that the listener would correctly guess the song title from the list, the tappers predicted about 50 percent. The actual success rate was around 3 percent. This discrepancy suggests that the tappers were unable to quarantine their own experience of the saliency of the song in order to come up with a reasonable prediction.

The egocentric bias also reveals itself in cases involving personal preferences rather than privileged knowledge. Van Boven, Dunning, and Loewenstein (2000), for example, conducted a study in which subjects were divided into two groups, buyers and sellers. The sellers were given coffee mugs and asked to estimate the highest price that the average buyer would most likely be willing to pay for the mug. The buyers, on the other hand, were not given but merely shown the mug, and then asked to estimate the lowest price that the average owner would likely accept for the mug. The result was that the sellers' estimates were much higher than those of the buyers, which can be interpreted as suggesting that they had as owners developed a preference for the mug and then projected that onto the buyers (this phenomenon is referred to as the "endowment effect"). I will merely note here that I am not fully persuaded of this interpretation, since there could also be other explanations of the discrepancy. The sellers could, for example, be biased by an assumption that since buyers are obviously in the market for mugs, they have a desire to buy a mug, while the buyers could be thinking that the sellers obviously do not their mugs anymore and thus do not value them very highly. I other words, there is an asymmetry in the analysis of the other group's motives since the other group in fact does have other motives, and this could result in the same discrepancy even without invoking an egocentric bias. Nevertheless, the egocentric bias remains a plausible explanation, and would be strengthened in conjunction with other, similar empirical findings.

Another study of the egocentric bias involving personal preferences was conducted by Ross, Greene, and House (1977). They asked subjects to wear a large sign reading "Eat at Joe's" for a half-hour. Subjects who agreed to do so also predicted that 62% of their peers would agree, whereas subjects who refused predicted that only 33% of their peers would agree to the request. Hence, subjects tended to expect others to act as they would act.

The third category of evidence for an egocentric bias is involves what appears to be a projection of one's own feelings onto others. Van Boven and Loewenstein (2003) related to their subjects a story about hikers lost in the woods and then asked what the hikers would likely be feeling. Subjects who had just engaged in physical exercise tended to predict that the hikers would be warm and thirsty; subjects who had not eaten recently predicted the hikers would be hungry, etc. Van Boven and Loewenstein interpret their findings as supporting the hypothesis that people predict other people's feelings by imagining how they themselves would feel in their situation.

In all these examples of egocentric biasing, what is going on is quite similar to what is going on when children fail false belief tasks. In either sort of case, people are (arguably) using themselves as a model for predicting and/or understanding others' behavior in an inappropriate way, failing to recognize relevant differences between themselves and others. And this is of course just the kind of mistake that simulation theory would predict. Where simulation theory runs into trouble (at least prima facie) is in explaining why we do not *always* make this mistake. This is not to say, though, that false belief tasks are inexplicable for simulation theory. Gordon in fact regards them as empirically *supporting* simulation theory, as we shall see in the discussion of Gordon.

Tomasello (1999) also thinks that simulation theory is better equipped than theory theory to account for the emergence of false-belief ascription around 3.5 years, and the reason brings us to the second kind of psychological argument in favor of ST: having experienced a situation themselves helps children to correctly predict how others will act in the same kind of situation. Tomasello mentions a study conducted by Perner and Lopez (1997), which found that young children "were better at predicting what another person would see in a particular situation if they had actually been in that situation first themselves."[39] His proposal is that the wealth of experience in discourse interaction that children have accumulated by this age helps them to take the perspective of others whose knowledge differs from their own. That is, they learn to re-phrase utterances that have not been correctly understood, to inform people of things that they apparently do not know, to engage in arguments with others whose perspectives on and/or evaluation of situations differ from their own, etc. In order to interact in these ways in such situations, they have to take others' perspectives, i.e. simulate them. Tomasello regards false belief ascription as a relatively advanced form of perspective-taking that results from this kind of practice made possible by language acquisition.

In support of this correlation with language, Tomasello cites a study by Appleton and Reddy (1996), in which children's performances on false belief tests improved when whey were engaged in discourse about the relevant mental acts. He also mentions a study done by Russell et al (1999), which showed that deaf children in general (most of whom have hearing parents, and who therefore do not have many opportunities to engage in extended discourse) do poorly on false belief tests. This is significant when compared to another study in which deaf children with deaf parents were found to perform normally on false-belief tests.

Another intriguing study that shows how first-experiences can be used as resources for children in getting a grasp of others' behavior was reported in 2004 by Sommerville, Woodward and Needham. Three-month-old infants were allowed to watch an adult grasping one of two objects that were located in between the adult and the child, both of which were too large and heavy for the children to be able to manipulate them (a ball and a teddy bear). After a while, the two objects were switched. In half of the trials the adults continued to grasp toward the same location, where there was now a different object. In the other half of the trials, the adults grasped for the same object, now in a different location. In the control case, the children looked longer when the adults grasped toward new locations than when they grasped toward new objects. In the test case, however, the children had been prepared by allowing them to wear special gloves that enabled them to move the objects around playfully. Having been prepared in this manner, the children looked longer when the adults' grasping actions were directed at new objects than when they changed target locations. Sommerville et al (2004) speculate that "agentive experience may contribute to the construction of goal-centered action representation during infancy."

Perner remains skeptical of this interpretation, however, noting that the agentive experiences that the children were able to have with the help of the gloves may have been

[39] This is Tomasello's gloss: Tomasello, M., 1999, p. 175.

only *indirectly* efficacious. The idea is that these experiences provided the children with motivation to observe the object manipulations more closely, since they themselves were participating in them, and gave them an occasion to notice a causal and therefore interesting connection between events – movement of the hand and movement of an object – but from the observer's perspective.

4.2.2.2 Physiological evidence

Motor resonance

Goldman was among the first to notice that work on mirror neurons provides empirical support for ST, and collaborated with Gallese on the influential paper that got the discussion of ST and mirror neurons going (Goldman and Gallese 1999). Since this is the topic of chapter 6, I do not want to go into it here, but the rough idea is the following. There are neurons in the motor system that are active when one is performing an action as well as when one observes someone else performing the same action. This suggests that understanding their action involves simulating it. At any rate, this is a prediciton that ST would make and that mirror neuron research appears to corroborate, whereas TT would make no such prediction. Feel free to jump ahead to chapter 6 and read a couple of pages before returning here if you like.

Paired deficits in experiencing and recognizing emotions

Another sort of empirical finding rightly emphasized by Goldman picks out a correlation between two deficits: namely emotional experience and emotion recognition. Ralph Adolphs and Antonio Damasio investigated a patient who has since birth had severe bilateral damage to her amygdala, which is involved in the experience of fear – including fear-based conditioning and the storage of fear-based emotional memories. According to Damasio 1999, she exhibits no social inhibitions, approaching new people and situation with a "predominately, indeed excessively, positive attitude."(66) She does not experience fear when shown clips form frightening movies, nor does she exhibit any of the typical physiological changes associated with fear. She recognizes intellectually when fear would be appropriate, but does not experience it herself. Interestingly, Adolphs et al also found that this patient was well below average in her ability to identify photographs of faces expressing fear. Sprengelmeyer et al (1999) found a similar correlation involving a patient whose hobbies included jaguar hunting while hanging on a rope from a helicopter in Siberia. This daredevil patient was also very poor at recognizing fearful facial expressions. More findings of a similar nature are reported in Lawrence and Calder (2004).

There are also some studies on psychopaths revealing the same link between deficits in experience and recognition of fear. Blair et al (2004) found that a group of psychopaths, who had reduced amygdaloid volume and also reduced amygdaloid activation during emotional memory tasks relative to a normal comparison group, were very poor at picking out facial expressions of fear.

4.2.3 Criticism

Goldman's critique of functionalism should certainly give us pause. It gives us good reasons to expect that some kind of direct method(s) for self-ascribing should be incorporated. This is intuitively plausible anyway, as a well-known joke about the two behaviorists illustrates: One behaviorist asks the other, "How am I?" And the other responds: "Fine, and myself?" On the other hand, Goldman's alternative grants qualia an implausibly central role. But it is not clear that his introspectionist account of mental concepts is a good alternative. It is not too terribly plausible to suppose that for all different mental states there is a different qualitative feel. Luckily, we do not have to follow Goldman in this respect. We can, instead, combine functional aspects with some more direct methods of categorization. Furthermore, we can be more clear about what notions of qualia or introspection make most sense and would best accord with other psychological and neuroscientific data. We will see later on, especially in the closing chapter, that Goldman's "introspection" – in particular as he has been speaking of it in recent years (e.g. Goldman 2006) - admits of a deflationary interpretation that fits well with psychological work on metacognition.

4.3 Robert Gordon

One of the other main proponents of ST, Robert Gordon, departs more resolutely from the TT than Goldman does insofar as he denies that either psychological laws or mental concepts are generally employed in folk psychological prediction. Accordingly, he calls his version of ST "radical", and points to three elements of other versions of ST that he rejects:

1. an analogical inference from oneself to others
2. premised on introspectively based ascriptions of mental states to oneself,

3. requiring prior possession of the concepts of the mental states ascribed (Gordon, 1995, 53)

So, obviously, his version of ST turns out to be quite different from rival versions. I will show in the following what he has against these three elements and how he intends to do without them.

4.3.1 Default ascriptions of knowledge

Gordon originally developed his theory in the context of an account of emotions presented in a 1987 book. His starting point for thinking about folk psychology in that context was his observation of what he calls the prevalent "factivity in our emotion concepts" (Gordon, 1987). What he means by this is that strong epistemic attribution of knowledge, rather than mere belief, is implicit in standard cases in which we explain people's emotions by giving reasons for those emotions. So, for example, we generally say "Smith is upset because Dewey won the election" – which implies that Dewey won the election and that Smith *knows* that Dewey won – and not "Smith is upset because he believes that Dewey won the election" (Gordon, 1987, 128). Only if we want to signal explicitly that we do not share Smith's beliefs do we say things like "Smith is upset because *he believes that* Martian spaceships are circling the Earth." This, Gordon urges, should be a surprising fact, since knowledge is epistemically more demanding than belief and requires additional evidence that go beyond belief. So why should attributions of knowledge be the default?

Clearly, the observation applies not only to attributions of reasons for emotions but also to attributions of epistemic states to explain all kinds of behavior (e.g. "Smith is tearing up his passport because Dewey won", "Smith is packing his swim trunks because it is hot in Bermuda"). Accordingly, Gordon's account is meant to apply broadly to ascriptions of epistemic states in predicting and explaining other people's behavior as well as their emotions. And since Gordon went on from these reflections to develop a full-blown account of folk psychology, we will of course have to see how he extends his account of belief-ascription to desire-ascription. But the basic idea is already present in the treatment of beliefs, so I will begin by focusing on this.

What one is doing in making such attributions, essentially, is simplifying matters by leaving out various complexities. One ignores issues like whether the other person really has access to all the facts that one is aware of, whether the person is justified in taking something for a fact (i.e. has a justified belief), etc. Hence, although attributions of knowledge appear to a philosophical analysis to be more conceptually sophisticated than attributions of belief, the default attributions of knowledge under discussion in fact bespeak a *lack* of sophistication. Our everyday concept of belief is carved out of our everyday concept of knowledge by making allowances for differences between our assessment of the facts and that of other people (or of ourselves at a different time). Making default attributions of knowledge therefore amounts to a failure to make allowances for the differences between one's own epistemic states and those of the other person; one is simply falling back on one's own assessment of the facts, and using this assessment to explain the other person's behavior in just the same way that one would use them to explain one's own behavior.

So let's consider for a moment how we invoke the facts in explaining our own emotions and actions. In one's own case (in explaining one's own behavior or emotion), the standard formulation also leaves out the reference to a belief. One does not say "I am Xing because I believe that p" but, rather, "I am Xing because p". Here, pointing at the facts rather than at a belief about the facts normally[40] makes no difference, since the assessment that one makes of the facts is the very same process by which one forms a belief. Referring to the facts from your perspective is equivalent to stating your belief about the facts.

The basic idea upon which Gordon is building here is the ascent routine devised by Gareth Evans (1982). Evans notes that whenever you are in a position to assert p, you are also in a position to assert you believe that p (Evans 1982, 225). Reliable self-ascriptions of belief, then, are not hard to come by. It is merely a matter of substituting one linguistic expression for another. You do not even need to master the concept of belief. Mastering the concept of belief would also require you to be able to ascribe beliefs to others (namely, beliefs you share as well as ones you know to be false, ones you believe to be false, etc.) and also to yourself at different points in time.

Of course, as I have already pointed out, we usually do not make the linguistic substitution suggested by the ascent routine. Doing so would not add any information in normal cases. So it is no surprise that the default method in normal cases is to explain our own emotions and actions by appealing to the facts, and not to our beliefs. So what about second- and third-person ascriptions? If Gordon is right that in these cases we fall back on an egocentric default method, then we would expect the same pattern in these cases as in first-person cases, namely, that we refer to facts and not to beliefs. And this is in fact the phenomenon that Gordon set out to explain.

[40] When talking or thinking about our own beliefs at different points in time, our present assessment of facts is of course not equivalent to the belief, since our beliefs change when we learn or forget information.

But, although this amounts to saying that we refrain from making the linguistic substitution suggested by the ascent routine in normal first-person cases since it would be redundant, and that in second-person cases we also refrain from making the substitution, since what we are doing is implicitly falling back on the first-person method, there is another sense in which what we are doing when we make a default-attribution of knowledge to someone else is analogous to using Evans' ascent routine. Let me explain.

4.3.2 The ascent routine

With the ascent routine, we move form an assessment of the facts to something that sounds like expression of a meta-belief via the application of a belief concept ("I believe that p"). Of course – and this is the point of the ascent routine, we do not really need a full concept of belief in order to do this. Gordon's key idea is to extend this to second- and third-person attributions. If we simulate (pretend to be, imagine being) the other person and assert p on their behalf, then we could also ascend to the assertion "I believe that p" on their behalf. Now we have something that sounds like a metabelief about the other person's belief. We could then end the simulation and retain the metabelief. Now we have done something that sounds like attributing a belief to them.

As I have said, the ability to make such an ascent does not constitute complete mastery of the concept of belief. This is just what Gordon wants: a construal of ST that avoids the need for introspection of mental states using mental concepts, and inference by analogy to the other person's mental states. Gordon formulates his solution by saying that ascribing a belief that p to someone amounts to asserting p within the context of a simulation of that person. Since extracting from the simulation a metabelief about the other person's belief does not require a full mastery of the concept of a belief or introspective access any more than extracting a metabelief about one's own belief via the ascent routine does, Gordon can say that ascriptions of beliefs on his version of ST do not involve the applying the concept of a belief to identify an internal state to which one has introspective access.

Since these ascriptions lack the sophistication of a complete concept of belief, they should be reliable only when these additional conceptual refinements are irrelevant. In first-person cases, though, we sometimes do add the qualification "I believe that". In other words, we depart from the default method in order to signal our uncertainty about the facts. In order to do so, we need a more complete mastery of the concept of a belief than we need when we have no reason to doubt our assessment of the facts. And so it is in second- or third-person attributions: we sometimes should distinguish between our assessment of the facts and the other person's. In these cases, we should depart from the default method, and in order to do so, we need a more complete mastery of the concept of a belief. So using Gordon's simulation procedure should be reliable whenever our own assessment of the facts corresponds to the other person's assessment, but not when the other person differs from us.

4.3.3 False beliefs and pretence play

It is on this basis that Gordon interprets false belief tasks as supporting his version of ST. The idea is that since understanding that other people can have false beliefs requires additional sophistication beyond just using the default method (i.e. of checking the relevant facts in the world and assuming that the other person's behavior is appropriate to those facts), children should first be able to understand true beliefs and only later attain the sophistication required in order to understand false beliefs. He argues that TT should not predict a different developmental timetable for true and false beliefs. If understanding beliefs involves internalizing a system of laws circumscribing the concept of belief, then there is no obvious reason why true beliefs should be any easier to understand than false beliefs. Once you have interalized the system, it should be equally applicable to true and false beliefs (1995, 70). I think Gordon is clearly right that the sequence fits his account well. And he may be right that TT explanations of this timetable are a bit ad hoc, but, as we have seen, there certainly is no dearth of TT explanations of this timetable. And even if they are a bit ad hoc, some of them are nevertheless pretty well worked-out.

At any rate, if we grant that the earlier understanding of true beliefs fits well with Gordon's account, the question still arises just what children learn between 3 and 5 that enables them to understand false beliefs. One option would be to say that they simulate up until then, but then learn that simulating is limited in that others can have divergent beliefs (as well as other divergent attitudes), and then they start to use a more theoretical apparatus at least in cases where the other is relevantly different from them. Goldman acknowledges that some sort of theoretical approach may often be used in such cases (Goldman 1995, 2006). But Gordon wants to avoid such a concession as much as possible[41]. He thinks that simulating somebody with a false belief is simply a more complex imaginative achievement, but that it rests on essentially the same procedure as simulating somebody with a true belief. In the case of the false belief, you would have to simulate the person's perspective over the course of an extended time period, i.e. from when they acquired whatever informaiton they have about the

[41] He would not totally rule out the possibility that some theorizing is sometimes used, but certainly he wants to avoid it for central cases such as false beliefs.

relevant situation up to the present. So for example, with the original Perner-Wimmer experiment, the child has to simulate the other actor seeing the chocolate bar placed in the drawer, then leaving the room, then coming back. She therefore has to simulate a somewhat longer sequence of experiences than in straightforward cases of true beliefs.

In support of this proposal, Gordon points to the correlation between failing to understand false beliefs and failing to engage in pretence play which is observed in autists (Gordon 1995, 70)[42]. He thinks it obvious that anyone who does not engage in pretence play would find it difficult to understand that others may have false beliefs since pretence play involves the same sort of procedure as simulation. He also notes that autistic children are at least as proficient as normal children at understanding mechanical operations. This fact, he asserts, is a "distinct blow to any functionalists who might think mastery of the concept of belief to consist in the acquisition of a theory of the functional organization of a mechanism." (1995, 70).

4.3.4 An appeal to common sense

Gordon also gives a sort of phenomenological argument for his version of ST.[43] If someone asks you whether you believe that Neptune has rings, how do you go about answering the question? Against TT, Gordon urges that you probably do not examine your recent behavior and weigh various hypotheses about what you believe. Contra Goldman, you probably do introspectively search for a "telltale feeling or other experiential mark of belief." (Gordon, 1995, 59) What do you do? You probably just reinterpret the question as "Does Neptune have rings?" If you are prepared to say yes to that, then say, "Yes, I believe it does have rings," adding the belief bit just because that was the way the question was phrased. There is no need for any additional evidence in order to make your answer into a statement about your belief. As Gordon says, this route to self-ascription via the ascent routine presents a non-

[42] I already mentioned this in the discussion of Alan Leslie's version of TT. Cf. Baron-Cohen, Leslie and Frith 1985.

[43] For some reason, a lot of people seem to find the word "phenomenology" cool this week, although it was certainly not so cool in analytic philosophy last week. So I invite you to pop open a Red Bull and contemplate this "phenomenological argument". Anyway, if you are not as trendy as I am, you can ditch the phenomenology crap and consider this argument an appeal to common sense or plausibility.

recognitional account that counters Goldman (1993)'s assertion that the only alternative to a (in his view highly implausible) functionalist account of self-ascription must involve introspection.

The same goes for desires. Which is good, because I promised a few pages ago to show how Gordon extends the account to desires. Since Gordon is responding to Goldman`s 1993 paper in favor of introspection in self-ascription, he takes up Goldman's own example of a desire – namely, the desire for a banana. On the face of it, the ascent routine seems not to apply here, at least not readily. The reason is that in the case of beliefs (e.g. "I believe that the banana is ripe"), you can get rid of the belief part and have a normal assertion. The ascent routine is a way to get from the normal assertion to the self-ascription. But what happens if you leave "I would like" out of "I would like a banana"? "Banana" is not even a sentence to be proud of, let alone the normal way of communicating a desire. The fact is, though, that this is an irrelevant little linguistic detail. The right conclusion to be drawn is not that we do not use the ascent routine here at all, but, rather, that the structure of the language demands that we make the ascent routine almost all the time (in standard cases anyway) when stating our desires. Just imagine how a kid learns to say "I want a banana" reliably when she in fact wants a banana. She starts out by just taking it or crying or whatever, and we teach her to say "I want a banana". Learning to substitute this new linguistic form for the taking or crying is equivalent to substituting "I believe that p" for "p". The child needs no new evidence. Nor does she have to have mastered the concepts of desire understood that "I" refers to herself. So Goldman is wrong to conclude that the child must have learned "some sort of (internal) cue by which she identifies the present state as a wanting." (Goldman, 1993, p3.65)

Note that Gordon is rather close to Josef Perner with respect to this difference vis-à-vis Goldman. This may be a bit surprising since Perner is a theory theorist. But, as we will see later on, Perner is actually not averse to introducing some simulationst components, and when he speaks favorably of ST, it is generally Gordon he has in mind. Both Perner and Gordon think the concept of a belief is an adjunct to folk psychology that is only used in special cases. They agree that folk psychology primarily uses situational information, not psychological information, to interpret other agents. Gordon points out that if children do not have a concept of belief until about the age of 4 years, then they must be using some other resources until then. It is sensible to think that they continue to use those other resources once they take on the concept of a belief.

4.3.5 Criticism

There is one obvious objection that one could make to Gordon's account. He says that simulations are more complicated when the target person is relevantly different from the simulator, as illustrated in the case of false beliefs. It seems that appreciated the relevant differences might require mental concepts. How else could we identify the relevant differences in order to set up the simulation properly? I noted earlier that he would spell out this idea of complexity, at least for the false belif task, by talking about how we need to simulate a whole series of someon's experiences rather than just simulating their experience at a given time. But how does the simulator know whether to simulate such a series of experiences or just the present experience of the target person? It looks like she needs already to understand whether the person has different beleifs in order to make a decision about which approach to take, thus presupposing an understanding of mental states! For this reason, I do not think Gordon can really get away with denying a role for mental concepts in such cases. It seems to me more promising to admit that mental concepts are involved but to give an account of them that avoids the intellectual baggage of TT. This, of course, is what Goldman tries to do. I am not fully satisfied with the result Goldman winds up with, but I do not agree with Gordon that doing something along these lines can be avoided altogether.

.

4.4 Jane Heal

In an article published in 1986, the philosopher Jane Heal set out one of the earliest formulations of simulation theory. In point of fact, the name that she initially proposed for the position was not "simulation" but "replication". Nevertheless, it clearly constitutes a version of ST and has subsequently been so classified by the participants in the discussion (herself included). I will therefore refer to it in the following as a version of ST. Similarly, in her earliest work on folk psychology, Heal refers to the opposing position not as theory theory but as functionalism. This is due to the simple fact that theory theory was the implicit consensus view in philosophy of mind, cognitive science and psychology that grew out of functionalism and did not have a specific name at all until simulation theory arose as an alternative in the mid-1980s. Later on she, like most others, explicitly links theory theory to functionalism. In Heal's version of the theory, the rationality at the core of folk psychology plays a particularly prominent role, as we shall see.

4.4.1 Critique of functionalism

Like Goldman, she develops and defends her position by means of a critique of functionalism. She considers a functionalist account of everyday psychological concepts to be implausible, namely because of the "holism of the mental"[44], as she puts it (with a glance at Quine (1960) and Davidson (1970). In other words, on a functionalist account, psychological states are defined in terms of each other (as well as via their links to perception and behavior). But this makes things quite complicated, since there is no limit upon how many other psychological states might be relevant to the definition of any one particular psychological state (2003, 12). She does not deny that a theoretical account is possible, but she points out that since we have not yet found an adequate one in cognitive science, it is implausible to suppose that we implicitly master such a theory. We should therefore be reluctant to adopt the functionalist account unless there is no alternative. But, of course, she thinks that the "replicating strategy" which she is sketching is "a real and conceptually economical alternative which avoids the need to credit ourselves with knowledge of complex theories about each other" (2003, 14)

She formulates her basic conception of the replication strategy, i.e. of simulation, as follows:

> I can think about the world. I do so in the interests of taking my own decisions and forming my own opinions. The future is complex and unclear. In order to deal with it I need to and can envisage possible but perhaps non-actual states of affairs. I can imagine how my tastes, aims and opinions might change and work out what would be sensible to do or believe in the circumstances. My ability to do these things makes possible a certain sort of understanding of other people. I can harness all my complex theoretical knowledge about the world and my ability to imagine in order to yield an insight into other people without any further elaborate theorizing about them. Only one simple assumption is needed: that they are like me in being thinkers, that they possess the same fundamental cognitive capacities and propensities that I do… (Heal 2003, 13)
>
> Suppose I am interested in predicting someone's action. What I endeavour to do is to replicate or re-create his thinking. I place myself in what I take to be his initial state by imagining the world as it would appear from his point of view and I then deliberate, reason and reflect to see what decision emerges. (Heal, 2003, 13-14)

There are a number of points here that we will want to address. Firstly, she, like Gordon, emphasizes that understanding other people involves first and foremost thinking about the world, not about their mental states. In subsequent writings, as we will see, she focuses increasingly upon the need for a simulationist account of how we think about the content of others' thoughts – i.e. rather than thinking about them thinking about the content of their thoughts, we simply think about the content (This will become clearer below). Secondly,

[44] Heal 2003, p.12

Heal speaks here of reasoning (and deliberating and reflecting). This is closely related to the first point, but needs to be picked out and treated on its own simply because Heal focuses increasingly on rationality, arguing that rationality is in particular susceptible to a simulationist treatment. Thirdly, she concedes the need for the "simple assumption" that others are "like me". This is a qualification of the first point. She departs from Gordon in that she thinks a judgment using mental concepts must be made in order to exploit a simulation to the end of understanding someone's behavior. Fourthly, there is an emphasis on counterfactual abilities (e.g. envisaging possible states of affairs, imagining how my tastes, aims and opinions might change), which suggests a connection with action planning that is worth expanding upon. I would like to discuss each of these points in turn.

4.4.2 Replicating content

As for the first point, Heal herself anticipates a line of criticism that could be levied against her. Specifically, the criticism is that we may need a functionalist theory in order to identify what state someone else is in before we can re-create their perspective and start the simulation procedure. In other words, if the simulation involves thinking about the world and not about others' minds, we still need to know what features of the world are relevant to the other person at the moment, what beliefs they hold about the world, and what goals they are pursuing. All this seems to require thinking about their minds. And if Heal's theory cannot tell us how this part works, then we are left with no alternative to functionalism. Or, at least, she has not given us one. Heal counters this with two lines of defence.

The first, which she herself notes is less important, is to stipulate that instead of falling back on functionalism, we "may propose instead some more direct model of how we come to knowledge of others' feelings" (Heal 2003, 15). Although it is surely unsatisfactory that she does not spell out how this more direct model would work and her "defence" therefore smells a bit like question-begging, the idea is not crazy. We will see in later chapters that there are theoretical and empirical resources available for developing this idea, but for now I will turn to the line of defence that she herself considers more important.

Apart from perceiving others' mental states somehow directly, the other possibility is to argue that in identifying what features of the world are relevant to the person whose behavior I am interpreting, I am still focusing my gaze upon the world and not on their mind, albeit from another perspective. Heal speaks here of a process of "recentring the world in imagination" (Heal 2003, 15). She illustrates this with the example of visual occlusion. I need some understanding of what one can perceive from where (e.g. that one cannot see through solid objects) in order to make the appropriate adjustments and figure out how the world looks from the different spatial perspective of someone else, but I do not need to know anything about how visual perception proceeds in the brain. Similarly, in interpreting others' behavior, I draw upon knowledge about their personalities, habits, obligations, etc., but I may be able to do this without employing mental concepts in the functionalist sense. Unfortunately, Heal does not spell out how this is supposed to work with, say, false beliefs or contrasting desires. Presumably she has in mind something like behaviorist-like ascriptions of dispositions to act in certain ways in certain situations. This would bring her closer to Gordon; I will compare the two of them explicitly further below.

I have noted that these lines of defence are not too well articulated. That may well have to do with the fact that Heal's interest in fact turns out to lie not in ascribing mental states to others but in predicting how people will act on the basis of mental states we have already ascribed to them, and on what further mental states they will judge to follow from present mental states. This comes out to some extent as early as the 1986 article, and in subsequent articles, Heal goes so far as to concede an important role to theoretical elements in working out what mental states people are in on the basis of their situations (Heal 1995, 50).

With respect to thinking about what actions or thoughts follow from thoughts that we already ascribe to others, she has presented some interesting arguments. One of them rests on the relevance of the content of someone's thoughts for predicting what future thoughts or actions those thoughts will yield – a point which, she submits, theory theory cannot adequately account for. Theory theory, according to Heal, conceives of people as employing general principles and making general assumptions about how other people think, i.e. that they think in accordance with modus ponens or that they will the means to the ends they have chosen (Heal, 1995, 35). But, Heal argues, the actual content of people's thoughts is always decisive for what subsequent thoughts or actions they will arrive at on the basis of those thoughts. If this were not so, then the putative folk psychological theory would present a shortcut to thinking in terms of content, and I would be better off employing it rather than thinking in terms of content even in predicting my own decisions. Heal concludes that this is silly: "Thinking about thinking cannot be easier than the first-level thinking itself" (Heal, 1995, 36).

She illustrates this point by reflecting that if one learns about a given subject matter, say quantum physics, one tends to automatically gain an ability to predict how other people who are knowledgeable about the same subject matter will respond to certain questions. One does not need a psychological theory. Anyway, even if one did employ a psychological theory, it would have to include all the representations of the subject matter in question, thus making it highly uneconomical. As Heal puts it:

> …whenever a person comes to be able to entertain a new kind of content, i.e. acquires a new concept, then a psychological theory for that person will need to be enriched by statements about the distinctive patterns of connection imported by that content. (Heal, 1995, 39)

How generally does this argument really apply, though? There does seem to be a distinction between cases where the subject matter is decisive and cases where contingent psychological factors play a role. Presumably, thinking about quantum physics (if I were able to do so profitably) would be the best way to predict how a quantum physicist would respond to a particular question about quantum physics. But how to go about predicting how my illiterate cousin Jed would respond to the same question? I happen to do enough about Jed to know that he would get defensive and bop me over the head. Getting that prediction right involves knowledge that may or may not be called theoretical, but it has nothing to do with the content at hand. Most cases fall somewhere in between these two extremes (Jed and the quantum physicist) with respect to how relevant content is. Heal seems to be focusing on cases at one extreme range of that spectrum. This is of course a shortcoming. But she would surely reply that the kinds of psychological principles the theory theory would involve are still not decisive in most cases, even where I do have to invoke facts about the other person rather than just thinking about the content of her thoughts. So even if the cases she picks out are not typical, theory theory does not thereby win out.

Indeed, she urges that the kinds of generality and platitudes that are often involved in making sense of other people's behavior do not qualify as a psychological theory in any reasonably robust sense of the term theory. A theory, according to Heal, should enable its user:

> …(to) talk of the interrelation of symptoms and of the contexts in which each is important. She can not only predict what will happen in a particular case, but locate that case among other possible cases… (Heal, 2003, 47)

What she is driving at here is in fact an interesting point, namely that employing a theory requires placing particular cases in a more general context in order to attain predictions. Heal thinks that much of this work that we normally expect from a theory cannot be delivered by a theory in the case of psychology. The argument is worth taking seriously. It turns upon what she says is an "important fact about thinking" namely that "justification or epistemic status is a holistic notion" (2003b, 51). Just about any thought at all could potentially be relevant in a given situation. Heal gives a deliberately outlandish example to illustrate this point. For a doctor examining investigating whether a particular patient has measles, the fact that Henry VIII was Welsh would seem irrelevant. The doctor simply performs a standard test and concludes, say, that the patient does not have measles. But. What if the doctor happens also to know that there is a rare genetic constitution in which the disease runs a non-standard course and for which the test is unreliable. If the doctor happens to know that this genetic constitution is common among Welsh and that the patient claims to be a descendent of Henry VIII, then she might be inclined withhold her judgment until some additional tests have been performed. In all cases where my actions or utterances, like those of the doctor here, are intended to be justified, all kinds of weird facts like this are potentially relevant. Hence, if I am predicting the someone else's behavior, such as the doctor's, then all kinds of weird facts are potentially relevant to making the prediction. Predicting what thoughts somebody will consider relevant involves making a selection from the multitude of possible thoughts. Of course this is a bizarre example. But Heal maintains that it shows an important point: theory theorists are committed to the implausible claim that our everyday folk psychological theory must include a solution to the frame problem. It is much more plausible to suppose that we bypass this problem when predicting what others will consider relevant to their judgments, actions and utterances by employing the same mechanisms that we employ when deciding what we consider relevant.

4.4.3 Rationality

We have seen that Heal's interest in primarily upon rational/normative connections among thoughts. This is indeed the distinguishing feature of her version of ST, and her thoughts on rationality among her most interesting and challenging contributions. It may seem counterintuitive to claim that rationality is a strongpoint for ST, since simulation appears to present an alternative to explicit reasoning in the sense of rule-based manipulations of representations, and after all, most people are inclined to conceive of rationality as consisting in this sort of thing. But Heal's reasons for placing rationality at the center of her version of ST merit careful consideration.

The starting point for these reflections is the observation that an important feature of our folk psychology is the assumption that others will be rational. If we did not make this assumption, or if the assumption were not appropriate, we would not get very far predicting how other people act. So it seems correct that characterizing this assumption is an important task in accounting for our folk psychological abilities. Heal's claim is that this cannot be achieved by postulating that a formal theory of rationality is part of our folk psychological competence – namely, because there can be no formal theory of rationality. Why not?

There are a few different arguments that one can find in Heal's writings, and she seems to waver among them. One central point is that an appreciation of our own fallibility and an openness to correction are important elements of our conception of ourselves as rational thinkers. Rationality, according to Heal, does not guarantee the actual success of judgment in particular cases, since it is also possible to raise the question "Have I got this right?" (Heal 2003, 20). Whatever set of inference rules or judgment-forming procedures we might point to as constituting the formal apparatus that would constitute rationality, it is always possible to question some or even all of these formal rules. Debating which rules are justified is constitutive of rationality, and it is an open-ended project. Thus, no static theory could ever capture it.

Heal adds another argument which I find less convincing. She notes that proposing a theoretical account of rationality would then make the assertion that I myself am a rational thinker (i.e. that the formal apparatus of rationality is instantiated in me and guides my thoughts) into an empirical claim, which could potentially turn out to be false. She does not like this:

> But, notoriously, any attempted demonstration to me by myself that I am a non-thinker must be absurd because self-undermining. Hence, any account of what it is to be a thinker which seems to make such a demonstration possible must be at fault (Heal 2003, 21).

I do not think this works. What is needed in the context of folk psychology is not a complete theory of rationality in general, but a theory of what assumption of rationality normal folks make in everyday life about the rationality of other folks. It could turn out that this theory does not apply to me that well, or even that it does not apply to anyone perfectly. But the conclusion would not be that I am not rational or that no one is rational, but that I am not rational in the specific sense of the everyday theory of rationality that people employ. The theory might turn out to be a decent fit but to leave out some aspects of our overall conception of rationality.

Another point that Heal makes is that our judgments of rationality are underdetermined by any formal account (1996). We do not always know what follows from our beliefs; we sometimes have inconsistent beliefs; we have to make an effort to be consistent and are sometimes corrected, we exchange reasons to defend our inferences. I think that this argument is open to the same criticism as the last one. What would be of use in folk psychology would be a theory of the kind of rationality that really underlies people's thoughts and actions, not an idealized theory. So it could simple incorporate these departures from ideal rationality that Heal points out.

A somewhat stronger point that Heal makes (1996) is that the prediction of rational behavior in any given case concerns not just principles of rationality, but *the application or interpretation of these principles* in novel cases. This is a Wittgensteinian point: what we would need is a theory that explains people's interpretation of the theory of rationality; and this theory would in turn have to be interpreted as well, etc. To avoid the regress, we have to use our own intuition about what is rational. A weaker form of the same argument would be that we always have novel constellations of beliefs and desires, so if we are to use principles to decide what follows from them the danger is that we need new principles for every novel case, and such principles are no principles at all.

I think this guarantees a simulationist component at the core of rationality, as Heal has in mind. But, importantly, it does not rule out the addition of a presence of a theory-like apparatus the application of which involves some sort of simulation. We will see later on that Josef Perner accepts that such a simulationist component probably has to play a role. So this idea of Heal's is quite compatible with hybrid theories.

At any rate, Heal's conclusion is that if there is no theory of rationality, then there is surely no folk psychological theory that includes a theory of rationality.

> So a corollary of the non-existence of a formal account of rationality is the non-availability of that mode of characterizing thoughts which functionalism counts on – a mode imagined to be independent of our entertaining or rethinking those thoughts (Heal, 2003, 22).

In short, we have to assume that how we are thinking is rational and that the other person will coincide with us insofar as they think rationally.

This line of thought is quite close to Hilary Putnam's thoughts on the impossibility of an evolutionary epistemology that fully naturalizes rationality. It is worth taking a glance at Putnam's reflections, since it will help to see the strengths and weaknesses of this line of thought. Moreover, since Putnam is taking issue with Quine's evolutionary epistemology, a look at his work will help to see the connection with the conception of rationality arising within the parallelist tradition that we looked at in chatper 2. Putnam's position is built upon his assertion that there are two aspects of rationality that are in tension with each other, but which both have to be accommodated. As he puts it:

"(1) talk of what is 'right' and 'wrong' in any area only makes sense against the background of an *inherited tradition*; but (2) traditions themselves can be criticized."(234)

For any belief or any theory, it has to be judged according to criteria that we accept as rational. And these themselves are open to debate. Even the application of them (this is a Wittgensteinian point) is open to debate. And so Putnam dismisses epistemological foundationalism. This is just fine for me, and I want to take issue with other stages in his argument, so I will leave his critics to criticize him on this point.

His arguments against naturalizing reason or epistemology are not fully persuasive. But the conclusion he arrives at is very sensible, and I think can even be compatible with a naturalized epistemology. So I will sketch his argument briefly. His first step is to give a definition of reason that is to be naturalized: "Reason is a capacity we have for discovering truths" (230). A big problem arises quite quickly then: the best notion of truth that is available is rational assertability, but we cannot say that "reason is a capacity we have for discovering what is (or would be) rationally assertable." So the evolutionary epistemologist has to presuppose a realist notion of truth. Putnam does not like this option, and so he dismisses it very quickly. I don't like it either, so I will only briefly mention his argument either for the sake of making clear what he is getting at, and leave the interested or unsatisfied reader to think it over more thoroughly. Aside from saying that a realist notion of truth would be "simply to revive the whole failed enterprise of traditional metaphysics," he also draws upon some remarks made by Roderick Firth[45] to the effect that even a realist notion of truth would not help. The idea is that whatever truth may be, we have no way of identifying truths but to posit that whatever beliefs are currently rationally assertable are true.

Putnam then offers the naturalized epistemologist the option of substituting some other notion for that of truth, such as beliefs "which promote our survival" (230). His rejoinder to this move is to point out that lots of irrational (manifestly ill-founded, arbitrary) beliefs can promote survival. But again, he glides past this issue so quickly that one has the impression he has not really vanquished his foe. He gives the following example: Science could produce weapons that are used in a destructive war that wipes out the human race. Putnam concludes that science would then prove not to have been conducive to survival, and would therefore have to have been an irrational undertaking. But this is not fully convincing. One could still argue that the cognitive habits and the justificational criteria used in science were very

[45] Firth made this argument in his Presidential Address to the Eastern Division of the American Philosophical Association (29 December 1980), entitled "Epistemic merit, intrinsic and instrumental".

conducive to survival in the long course of human history even if in some isolated cases it failed. The successful track record overshadows the one glaring failure.

But, in addressing Quine's version of naturalized epistemology, Putnam comes to a conclusion that is quite interesting. He says that the only notion of truth that Quine, as a naturalist, would accept is based upon the idea that someone believes a sentence to be true if he or he would assert it. Sentences are true in a target language if they are true in the metalanguage or the language of the interpreter. So a rational method for arriving at such sentences is just a rational method of arriving at sentences that I would be inclined to assert (245). This is a simulationist notion at the basis of rationality. Indeed, this fits well with Quine's remarks about empathy grounding the principles of charity.

Of course, Putnam is right that all this is purely descriptive and does not amount to normative epistemology. But – and here I am repeating my criticism of Heal – this does not matter if what we are looking for is an account of rationality that applies to normal people in everyday life. Quine is of course looking for more than this, and perhaps Heal is as well, but the decisive point is that *she does not need to*. Her focus is not epistemology but psychology.

So, in summary, it seems possible – contra Heal – that there could in principle be a theory of rationality that would be applicable to people's behavior in everyday life, even if it is perhaps impossible, or at least at the moment unimaginable, how to theoretically capture the *entire* concept of rationality, including the entire normative dimension. Nevertheless, if one interprets her arguments in an empirical vein, the point is well taken that using our own assumptions about others' rationality (i.e. simulating their thought processes) is probably at least as good as and certainly more parsimonious than using an additional theory about rationality that would not be likely to contribute any additional explanatory benefit.

4.4.4 Ascription and the nature of mental concepts

In order to round out the account of psychological concepts, of course, Heal also has to say something about first-person ascriptions. Given her affinity for Wittgenstein, one would not expect her to go too far down the introspectionist path that Goldman, although we have seen that her version of ST does in certain respects come close to Goldman's.

What she says is that that first-person ascriptions also depend upon a replication – or simulation – process. "Thinking about my own thought is… in my own case as for others, replicating - that is, putting on a certain sort of performance, rather than being in possession of a certain kind of quasi-perceptual knowledge." (Heal, 2003, 26) It is not clear whether these remarks about the replication process are meant to apply just to the movement from mental states to subsequent mental states (rational thought, inference) and to actions (practical reasoning) or also to the first-person ascription of a mental state in the first place. Her focus on rationality/ cognitive competence would seem to imply the former, which would leave her with some kind of non-replicative account of ascription along the lines of Goldman.

But I do not think this is what she has in mind. For she goes on to say that ascription "might better be called re-expression than description" (26). The use of the term "expression" of course harkens back to Wittgenstein, and suggests an interpretation that is closer to Gordon than to Goldman. And when she, in passing, wants to clarify what she means in speaking of the technique of replication, she writes: "namely, looking at the world."(27) The idea here, very roughly, is that one looks at the world, assesses the situation, and finds oneself inclined to react somehow or to feel a certain way or whatever. These responses are expressions rather than descriptions of one's mental states. Through enculturation, one observes others reacting the same way, making similar facial expressions, etc, and also making assertions about what they want or believe and how they feel, and one learns to correlate the utterances with one's one dispositions to act, feel, etc., and comes to use the utterances sometimes instead of or in addition to just action, feeling, etc. But the basis for doing so (for uttering statements reflecting one's mental states) is the same as the basis for acting or making particular facial expressions. The former, then, are nothing other than substitute expressions of one's mental states.

As for the difference between first- and third-person ascription, i.e. the question of privileged access, she asserts that first-person cases involve the same technique but do not involve the "complexities of re-centering", and are therefore easier. But she goes on to say that we have "no privileged position in connection with claims to understand it (my own thought), see what follows from it or the like." (Heal 27) Again, here, she seems to deny a privileged perspective upon thought processes leading from a mental state to other mental states and to behavior, but to affirm a privileged perspective upon what one's mental states are. But this, I think ,would be mistaken, for deriving consequences form one's mental states ("seeing what follows form it and the like") is part of fixing the content and thus achieving an ascription. Once again, then, I think we should avoid reading her as an introspectionist with regard to first-person ascription. We can make more sense of her remarks by taking her to mean that we are privileged with regard to our own minds in that we know what we presently perceive or have recently perceived, whereas we can acquire equivalent knowledge of others only to a limited degree and with the help of additional processes, such as imagining ourselves in their physical position, recalling their earlier physical position, asking questions and so forth. A limitation of this view is, of course, that mental states as theoretical entities become more efficient than complex imaginative processes when the situation becomes complex.

The upshot is that her conception of the relation between first- and third-person ascription is more or less diametrically opposed to that of Goldman. While he just bites the bullet (in fact, he does not seem even to notice the metallic flavor of his fare) and makes third-person ascription parasitic upon first-person ascription, and founds the latter upon introspection, Heal makes even first-person ascription depend upon replication.

4.4.5 Criticism

I think it is a weakness of Heal's theory that she does not engage much at all with empirical work. Indeed, she brushes aside the empirical work Goldman points to, saying that it is irrelevant since ST should be conceived as an arguing that understanding others' psychological states must *in principle* involve simulating those states (1995, 2003). Of course, some philosophers might consider this to be a strength, and might appreciate the a priori character of her reflections concerning rationality, since they constitute genuine philosophy rather than the sort of mixed, empirically motivated theorizing that most other participants in the debate engage in. I do not want to brush aside this attitude, and indeed I think her reflections on rationality are an important contribution, but I see no reason why one should not supplement them with empirically based theorizing.

The other point of criticism I want to reiterate is simply that her arguments against the possibility of a theory of rationality are not completely convincing, since they do not exclude the possibility of a theory that is good enough to apply to people's everyday behavior and make sense of it as rational, even if a complete theory of the entire domain of rationality, including scientific rationality, may well be unattainable. I think this second point of criticism is in fact connected with the first, since it reflects the a priori character of her argumentation. This misses a point that is crucial in the context of the discussion of folk psychology: it is not rationality in principle that needs to be accounted for but people's assumptions about others rationality, which is an empirical issue.

4.5 General Discussion

4.5.1 Contrasting the three accounts

As I have pointed out during the course of this chapter, there are important differences among the three main versions of ST that I have discussed. If we compare them to TT, Gordon's version is most radical in that he not only denies that psychological generalizations play a significant role in folk psychology but also that ascriptions of mental states using mental concepts play a significant role. I have already noted that, altough I think Gordon is right that our our knowledge of situations, facts about the world, and the like plays a key role, an adequate theory of folk psychology must also make more room for understanding of others mental states, especially insofar as other are relevantly different from us. Mental concepts like belief and desire are the obvious way to do this, but Gordon only tells us how the ascent routine enables us to do without full-fledgd mental concepts. This is not enough; we also need an account of mental concepts.

Goldman rejects only the first of the two components of TT (psychological generalizations), and carves out a substantial role for mental concepts. Throwing out psychological generalizations of course means that he cannot appeal to functional accounts of mental concepts and thus flirts with introspection. Since introspection is a problematic concept, this is a step that many people – including myself – would prefer to avoid. Another reason why genuine introspection is a bad idea, I think, is that we want to make sense of implicit simulation (since we patently are not always consciously doing running simulations), so introspection in the sense of looking inward and calssifying one's own mental states using concepts seems too sophisticated anyway. It may be reasonable to replace introspection with some sort of automatic unconscous monitoring of one's own mental states. Spelling out how this might work will be a task for a later chapter.

But I just want to note, with an eye to Jane Heal as well as to Piaget, that spelling out this idea would suit both of their programs quite well. Indeed, I think that substituting an implicit, automatic monitoring process for introspection could be said to revive the Fechnerian idea of an inner perspective in a deflationary way. The key idea would be that undergoing particular mental states and/or processes enables us to predict the way they would unfold if we were to act upon them, but without having any knowledge of how one came to that prediction. In other words, the prediction would be based upon procedural knowledge rather than on propositional knowledge. The predictions would work insofar as we are similar to others. And since rationality would be a core feature of that similarity, being a rational agent would enable one, as Heal has it, to construe others as rational without employong a theory of rationality.

4.5.2 Criticism

One problem with simulation theory is that it is simply quite diffuse. Different people mean different things when they talk about simulations. Some propose that in folk psychology we actually imagine ourselves in another person's situation in order to understand or to predict their behavior. I think it is plausible that we do this sometimes, but we do not seem to do it actively all the time. So what else could simulation mean in this context? Some are inclined to think in terms of implicit simulations, i.e. unconscious, automatic processes. Gordon (1995) endorses this idea, but is not clear about what it could mean. Goldman means to spell it out in terms of mirroring (e.g. mirror neurons, see chapter 6). Some, on the other hand (especially Heal), think of simulation as being in principle at the core of understanding psychological concepts. These are all very different ideas, and it does not seem promising to try to reconcile them all. Still, I think one can pick out a core insight of ST, even if the way in which this insight is spelled out differs among the varous versions: namely that *we undergo some of the same states and/or processes in explaining and predicting others' behavior as we undergo when we ourselves are deciding upon, planning and executing actions, and that this overlap obviates the need for extra, specifically folk psychological knowledge.*

Secondly, simulations alone cannot work insofar as we are relevantly different from others. So we need to identify the ways in which a target person differs from us in order to set the appropriate parameters. This arguably requires mastery of mental concepts defined functionally by their nomological relations among each other and to behavior, since our own beliefs, desires, habits etc. prima facie can't be of much use to us in figuring out to what extent other people have different beliefs, desires, habits, etc. So we are left with no alternative to inferring their beliefs and desires from their behavior and from what we know about their perceptions. I don't want to rule out the possibility that there are ways of dealing with this problem within a simulationist account[46], but most people agree at least that it presents a problem.

Thirdly, it is questionable whether ST is phenomenologically any more plausible than TT. If it seems strange to say that we ascribe others mental states and apply generalizations in everyday life, it may not seem any less strange to say that we constantly imagine ourselves in their shoes and simulate their actions, decision, feelings, etc.

[46] Gordon thinks it can be dealt with better within a simulationist account than a theory-theory account, Cf. Gordon 2003.

Chapter 5:
Comparing, Condemning or Combining Theories of Folk Psychology

Having presented both TT and ST in their multifarious versions, I would like in this chapter to look at efforts to decide the deabte between TT and ST, to replace them altogether, or to combine them. I will start (5.1) with efforts to decide between the two theories by weighing theoretical and empirical arguments. We will see that attempts in this vein have not been all that successful. This may be taken to suggest that the distinction between the two theories is not sharp, or that it is not productive to try to choose one of these two theories at all. Some people think – not only but also for this reason – that both theories should be abandoned or that the terms of the debate should be thoroughly re-thought. I will review a few such perspectives (5.2). Although I think many of the points raised by these authors are reasonable, I think it is worth trying to make use of the insights brought out by TT and ST, and also to try to to use the distinction between TT and ST productively, although I also conclude that the most productive path to follow is to work toward a hybrid theory. I will make some remarks about how such a hybridization could work, and and will discuss a couple of proposals that have been made for hybrid theories (5.3).

5.1 Comparing theories of folk psychology

It should be clear by now that there is no simple opposition between two theories of folk psychology but, rather, a considerable variety. Obviously, this makes it difficult to decide between/among them. Nevertheless, we can weigh the various theoretical and empirical points that favor one or the other theory (or, rather, one or the other version of one or the other theory), and in so doing, investigate which theory better embodies general theory-virtues such as theoretical simplicity, empirical scope, synthetic or cohesive power, empirical fruitfulness, etc.

Aside from these general virtues, we will also want to bear in mind which theory better accounts for such features of mentality as intentionality, semantic opacity, privileged access, mental causation and qualia. Assessing the theories with respect to these features has two aspects. Firstly, a theory of folk psychology should account empirically for how these features of our everyday mental concepts arise. In other words, it should answer questions like: what is it about mental concepts that they should have such features? How do these features fucntion in everyday explanations and understanding? These are empirical issues. It certainly does not count against an empirical theory of folk psychology if it winds up presenting a picture of the mind that involves philosophical difficulites if those philosophical difficulties really are inherent in our mental concepts. Secondly, it will be additionally attractive if a theory of folk psychology gives us a characterization of mental concepts that suggests novel ways to deal with the philosophical issues surrounding the features in question.

Although we will want to bear the variety of versions of TT and ST in mind, it will be helpful to start out with a baseline version of each theory that we can use to compare them. For TT, the baseline version has it that folk psychology rests upon the ascription of mental states such as beliefs and desires as causes of behavior, which are defined by generalizations concerning tying them to perception and behavior. The baseline version of ST holds that in everyday social cognition, we undergo some of the same states and processes as the people whose behavior or experiences we are seeking to explain or understand, and that this symmetry obviates the need for at least a significant portion of the psychological knowledge that TT postulates.

5.1.1 Empirical Considerations

I will start out by mentioning some early psychological studies aiming to decide between TT and ST empirically, and then move on to neuroscientific studies done in the past 10 years. It is an interesting fact that there is a sort of chronology here, the psychological studies having been done before the neuroscientific studies. I think the reason for this is that psychologists in the late 90’s gave up trying to decide between TT and ST, since the studies up to that point were all inconclusive, and it became clear that either side could account for almost any data. This suggested – and suggests – that the dichotomy between the two accounts is infelicitious. And indeed, as I have already mentioned, the trend has been toward hybridizing. On the other hand, once neuroscientists got involved in the discussion, they naturally began to look for neuroscientific ways of resolving the dispute - hence a fresh wave of neuroscientific studies in the past ten years. But let me start with just a couple of the earlier psychological studies.

5.1.1.1 Psychological Studies

In favor of ST, one could point to the various studies revealing an egocentric bias that were discussed in chapter 4 (4.2.2.1). Obviously, those studies speak in favo of ST because they suggest that people expect other people to act or judge the way they themselves would. Especially in cases where the way they act or judge is idiosyncratic, this suggests that they are using themselves as a model of others rather than appealing to generalizations.

On the TT side, Perner and Howe (1995) conducted what they conceived to be a sort of crucial experiment. They presented children with a story about two characters – John and Mary. John tells Mary that he will put a box of chocolates in either the bottom drawer or the top drawer. Mary goes to the library. John puts the chocolate in the top drawer and then goes to the park. Then, while John is in the park, his mother comes home and moves the chocolates to the bottom drawer. The children, who were all between about 5 and 6 and thus old enough to understand false beliefs, were then asked three questions:

Q1: Where does John think the chocolate is?
Q2: What if we go over to the park and ask John: "John, do you know where the chocolate is?" What will he say?
Q3: What is we go to the libraray and ask Mary: "Mary, does John know where the chocolate is? What will he say?
(Perner and Howe 1995, 164)

Perner and Howe claim that TT and ST predict different patterns of response. ST should predict that Q1 and Q2 should be answered equally well and better then Q3, since the very same act of simulation that would be used to answer Q1 would also yield an answer to Q2. But Q3 would require a more complex simulation (if simulation is used at all), since the child would have to simulate Mary simulating John. TT, on the other hand, predicts that Q1 will be easiest and that Q2 and Q3 will be equally hard. According to Perner and Howe, answering Q1 requires having a thought about a thought ("John believes that the chocolate is in the top drawer"), whereas Q2 and Q3 demand an additional level of complexity since the child has to think about somebody else's thought about a thought (for Q2: "John believes that he knows…" and for Q3: "Mary believes that John knows…")

The results roughly favored TT, as the children found Q1 easier than Q2 and Q3. Still, it is certainly not clear that Perner and Howe are justified in attributing the predictions to ST that they do attribute. Gordon (1995), for example, disputes this. He thinks it natural on a simulationist account that Q2 and Q3 should be more difficult than Q1. He reasons that the children may be influenced by their understanding that one should not claim to know something if one has reasons to doubt that it is correct, and since they themselves have reasons to doubt John's knowledge, they simulatively attribute these reasons to John or to Mary. Of course, Perner's argument that ST should predict that Q3 would be more difficult than Q2 still stands. So, although this is far from a decisive experiment, it should be counted as a victory for TT.

While we are talking about Perner, here is another empirical argument he advances against ST. The developmental timetable of children's understanding of their minds as representational media, and of the possibilities and pitfalls that this entails, is also reflected in their understanding of others' minds. That is, Perner argues that there is a parallel development of first-person and third-person knowledge of mental states. So, for example, he describes a set of experiments in which children either guess about or visually gain knowledge about the location of an object. In the cases where the children guess correctly, they should have answered that they had not known but merely guessed, and indeed children older than 4 were quite good. But younger children could not make correct use of the distinction (1991, 156-8). The relevant point is that their competency at applying this distinction to themselves develops simulataneously with their competency at applying it to others. And this is just one of numerous pieces of evidence for developmental synchrony with respect to understanding knowledge, i.e. distinguishing it from mere guesswork and from belief, recalling the sources of knowledge, etc. (Perner 1991, 270, Hogrefe, Wimmer and Perner 1986). Perner thinks this parallel development speaks against ST, which should predict that children's understanding of minds develops more quickly in first-person cases and then gets applied or projected onto others. I am not fully convinced that this parallelism really speaks against ST. It would certainly speak against ST if third-person competency were to emerge sooner, but parallel development leaves open the possibility that the emerging understanding of mental states and of their representational nature can more or less immediately be applied to others without any noticeable time lag. I see no reason why this should not be a plausible position to take.

Another psychological study favoring TT and discussed in Stich and Nichols (1995) has to do with so-called position effects. Presenting with products of equal worth, people will tend to choose the product on the right-hand side (Nisbett and Ross 1980). But, as Nichols and Stich point out, almost no one who hears about these experiences claims to have expected that result. This suggests that in forming an expectation of how others will behave in the experiment, people do not simulate them, because, if they did, they would presumably also be inclined to chose the object on the right and thus make the correct prediction for the other person. I do not find this all that compelling. At the most, all it shows is that there are influences upon decision-making that we fail to take into account of in running simulations. In other words, rather than showing that we do not simulate at all in this case (as Nichols and Stich suggest), it could be taken to suggest that our simulations are inadequate in some cases.

5.1.1.2 Michael Tomasello's Developmental and Comparative Approach

The developmental psychologist and comparative primatologist Michael Tomasello[47] argues that humans are distinct from other primates in regarding their con-specifics as intentional agents who pursue goals and employ strategies in doing so, i.e. in employing folk psychology. Tomasello gives some important developmental and comparative reasons for thinking that this folk psychology should be conceived along the lines of ST.

Just to pick out a few of the numerous bits of evidence for the basic thesis that humans are unique in emplyoing folk psychology, Tomasello points out numerous kinds of activity in which non-human primates do not engage. They don't point at objects, they don't hold objects up to show them to others, they don't lead others to places in order to show them objects, and they don't deliberately learn or teach new behaviors (Tomasello (1999). Moreover, they are unable to understand gestural or iconic hints about where objects are hidden.

They emulate human experimenters' actions (producing the same results), but do not imitate the strategies employed. That is to say, apes are good at learning to reproduce (emulate) events, i.e. effects upon the environment, but are not so interested in or adept at learning to employ the same technique or strategy as a human experimenter. For example, human children will imitate strategies even when they are inefficient, such as using a broom to get an object from the shelf when it could be done without the broom, or using their heads to turn on a light when their hands are free (unless there is an apparent reason why the adult used her head, such as having had her hands full) (Meltzoff 1988a). Apes do not do this. Tomasello's gloss is that they are not interested in the human's perspective upon what she is doing, or in sharing this perspective, but only in perceptible events. Also, humans recognize intentions in incomplete actions and imitate the result that they did not observe rather than the motions they did observe. Apes don't do this either.

With regard to the second thesis, namely that this folk psychological understanding of conspecifics should be conceived along the lines of ST, Tomasello points out that it is around nine months that many uniquely human behaviors begin to emerge. This is about the same time that children begin to employ various strategies to attain an end (e.g. moving pillows that obstruct their path to a toy) and to employ means to an end (e.g. pushing adults' hands toward objects so that the adults pick up the objects). Tomasello's theory is that this emerging sense of their own agency is a decisive resource that they use in interpreting others' behavior. Apes are less able to employ this sort of analogy from me to you and also less interested in dong so. So, what Tomasello gives us an argument that the developmental sequence of first- and third-person udnerstanding of agency (contra Perner) does indeed support ST.

5.1.1.3 Neuroscientific Studies

I would like to turn to some fairly recent neuroscientific studies that have been conducted with a mind to deciding between TT and ST. Apperly (2008) has published a review in which he discusses various studies and takes the position that they are informative but do not decide between ST and TT. Discussing Apperly's criticisms of the studies will provide a starting point for working out what kinds of empirical facts could decide between ST and TT.

Ramnani and Miall (2003) conducted a study comparing neural activation in predicting or interpreting others' behavior, on the one hand, and planning or executing one's own behavior (or predicting it in counterfactual situations or interpreting one's own past behavior). The idea here is that an overlap between these two types of cases would support ST, since ST would predict that groups of neurons involved in first-person cases are also involved in third-person cases, whereas TT would have to reason to predict this. The setup was a sort of game with two human participants and a computer. The participants were presented with colored shapes. The color determined which one or the other of them, or the computer, was to press a button. The shape determined which button. They were also instructed to monitor whether the other person and the computer pressed the correct buttons. FMRI was used to compare the neural activation in the premotor cortex when the participants were preparing to act with activation when they anticipated the other person's action. The authors found no overlap and concluded that their study corroborated TT.

Apperly is not convinced that the test addresses ToM skills at all, since the participants can correctly anticipate responses using "stimulus-response mappings"(271), i.e. without considering internal mental states such as beliefs and desires. But, contra Apperly, there are versions of ST (e.g. Gordon) according to which we do not commonly represent others' mental states in predicting/understanding their behavior. Such versions of ST would indeed predict an overlap here. In fact, the setup applies in general as a test between TT and ST if we allow for implicit ascriptions. After all, a simulationist should expect that whatever processes are involved in the participants decision and/ execution should be involved in predicting the other person's behavior. But if, as the results suggest, one in fact uses stimulus-response mappings in predicting the other person's behavior, the simulationist prediction is not borne out. Apperly is right that the test does not decide between ST and TT as accounts of how we predict and understand others' behavior in cases that involve more complex reasoning about mental states, or reasoning about mental states that diverge form ours. And of course, even if restricted to implicit ToM skills, it is just one test and does not decide the issue. But it should count as a victory for TT.

Another series of studies pursuing roughly the same strategy has been conducted by Grezes, Frith and Passingham (2004a, 2004b). Participants were videotaped while picking up boxes. In most case, they were correctly informed about whether the box was heavy or light, but in some cases they were misinformed. In the test phase, they watched videos of themselves and of other participants picking up the boxes and were asked to guess whether the person in the video had been informed correctly or misinformed, i.e. whether they had a true or a false belief about the weight of the box. As it turned out, there was activation bilaterally in the premotor cortex when the participants were judging themselves and when they were judging others. The authors argued that this common activation provides evidence for ST by suggesting that the participants were using the same circuits for modeling their own actions and those of others.

Apperly is unimpressed for basically the same reasons as he gives against the Ramnani and Miall (2003) study. He maintains that, although the participants were asked to formulate guesses about the person's belief states, thus introducing ToM content, the neural activation observed here does not necessarily have to do with this ToM content. It could well be that the neural activation reflects predictions about the person's movements on the basis of perceptual cues, but that reflections about belief states are not instrumental in making these calculations. This is true enough, but I would reply here again that the formation of expectations can still be said to involve implicit ascription, even if the explicit, verbalized ascription is, as Apperly contends, a separate matter altogether.

But Apperly raises a second point, which I take to be right on the money. Specifically, it is problematic that the first-person condition involves a video of oneself in the past rather than current first-person action. A theory theorist could explain the overlap by arguing that in observing oneself in such a case, one takes a theoretical stance just as one takes when observing others. So the study does not go very far in discriminating between ST and TT.

There is a study by Vogeley et al (2001) that has a similar structure, except that the first-person condition involved a hypothetical self rather than a past self. The participants heard stories in which the protagonist is either a fictional character or themselves, and were led to make judgments either about the mental states of the protagonist or about physical causality. For example, in some physical causality trials they were asked how it came about in the story that a burglar alarm was activated. Here, the right answer was a physical explanation. On the other hand, in trials targeting ToM skills, they were asked about how the protagonist would feel (e.g. after having been robbed). Interestingly, regions of the right anterior cingulate cortex and left temporopolar cortex were more activated for ToM judgments that for physical causality, irrespective of whether the protagonist was oneself or a fictional character. However, a region of the right prefontal cortex was more activated for the self-perspective ToM condition than for any other condition. The authors argue that the overlap supports ST, whereas the dissymmetry in the right prefrontal cortex supports TT, and conclude that a hybrid theory is most likely.

Apperly is not too terribly impressed, since the comparison does not involve current first-person action execution. He is surely right that the latter would be quite interesting to compare, but I do not share his utterly negative conclusion that the study says nothing about the ST/TT affair. The first-person condition in this setup is in fact crucially different from the setup in which the participants watch videos of themselves (and others). In the setup with the videos, they are not motivated to actually recall or imagine themselves carrying out the actions; they can very easily take the same detached perspective upon all the videos. In Vogeley et al (2001)'s study, on the other hand, the participants are actually encouraged to imagine themselves undergoing the events of the story, and thereby to take a first-person perspective upon the narrative. Moreover, there is in fact plenty of evidence that many of the same circuits are involved in imagining an action as in executing it[48]. In combination with such evidence, the first-person condition in Vogeley et al (2001)'s study could be regarded as probably overlapping with a first-person condition in which the self is not hypothetically but currently executing an action (or being involved in a story, as it were). Hence, the evidence conditionally supports the conclusion they draw, i.e. some overlap but also some asymmetric activation, hence a hybrid account.

The strategy employed in another kind of study that Apperly discusses is to check whether neural activation in third-person judgments is modulated by perceived similarity to the target person. The idea behind this strategy, obviously, is a bit different. To whit, the argument is that "simulation accounts of mental state attribution suggest that perceivers only use self-reflection as a strategy to predict the mental states of others when these individuals are in some way similar to self" (Mitchell et al. 2005, 1307, Cf. also Mitchell et al., 2006, Saxe and Wexler, 2005, Frith and Frith 2006). Hence, ST would predict that overlaps in

[48] I will not discuss this here, since we will be addressing it in more detail later (see chapter 6). This defence of Vogeley et al (2001)'s conclusion can be regarded as conditional upon that evidence.

neural activation between first- and third-person judgments should increase with the perception of similarity between oneself and the target person.

In summary, at least two points of criticism can be directed at these studies *as tests intended to discriminate between ST and TT*. Firstly, comparisons of neural activation between first-person and third-person cases should refer to first-person cases in which the participant is currently undergoing the relevant process (e.g. making the decision, performing the action, etc.). Contra Apperly, however, I think that experiments in which participants are encouraged to imagine themselves in counterfactual cases are indeed also worthwhile, much more so than experiments in which one observes videos of oneself and is not encouraged to imaginatively take an active role. Secondly, it is not clear that the tests really involve ToM skills. Apperly argues that the belief ascriptions involved here may amount to nothing more than interpreting the target's expectations on the basis of their motor behavior. I have pointed out that such a lower-level kind of ascription without mental concepts is indeed also at issue between TT and ST, and is in fact primarily at issue in Gordon's ST. So, yes, the force of the studies is limited, but it is not as limited as Apperly suggests.

There is a follow-up point that Apperly makes to the second objection, which is also worth thinking about. He notes that it is not only difficult to assess when someone has ascribed a belief, but also when someone *has* a belief. If we want to compare activation between first-person cases where someone acts on the basis of mental states such as beliefs with third-person cases where they ascribe these beliefs, then we need to have a way of determining when they have beliefs. But in the studies under discussion, there is room to doubt whether the participants have beliefs even in the first-person conditions. If they are told that the box is heavy in the Grezes, Frith and Passingham (2004a, 2004b) studies, for example, they may well suspect that the experimenter is trying to fool them, and thus either form the belief that the box is light or withhold belief. Moreover, Apperly argues, the implicit belief that is at work in formulating a motor plan for lifting the box may not qualify as a full-fledged belief. He does not argue this point sufficiently, simply noting that the folk psychology debate has been directed toward "personal-level beliefs and not at implicit, or subpersonal, motor beliefs". As I have noted, this it simply not true. But Apperly casually refers to a body of research that does indeed challenge the interpretation of motor beliefs as beliefs. Such "motor beliefs" – manifested by "the ability to perform accurate, visually guided, object-directed actions" – can doubly dissociate from "the ability to make judgments about the size, shape or orientation of objects on the basis of some incoming visual information" (Apperly 276). Apperly is right about this. I will not go into it in detail, but there are lots of studies involving people with various kinds of brain damage who are impaired either in their ability to recognize objects or shapes, or in their sensorimotor ability to manipulate the objects properly, but not both (Milner and Goodale, 1995; Jacob and Jeannerod, 2003). This data suggests a functional and anatomic distinction between the sensorimotor perceptual system and the object recognition/ conscious system. If outputs of the former system are not accessible to the latter or to other cognitive processes, then there is reason to deny them the status of beliefs. For if we are ascribing beliefs on the basis of behavior (including verbal reports), then some of the evidence will speak against ascribing a belief in these cases – to whit, the verbal report expressing ignorance of the object's shape. It seems we have to either deny them beliefs or ascribe contradictory beliefs.

Finally, I think it is fair to say that these studies are limited in that they compare third-person interpretation with prediction, recollection or counterfactual imagination of first-person action, rather than with actual first-person action. Another more interesting case to look at would be activation in predicting or interpreting others' behavior, on the one hand, and actually executing or planning actions on the other hand, rather than predicting or remembering one's own actions. MN research is a step in this direction, and we will discuss that in detail in chapter 6.

MN research, as mentioned in section 4.2.2.2 – and as I will discuss at length in cahtper 6 – provides empirical supprot for ST, because it suggests that at least some of one's own mechanisms for action planning and performance are used in interpreting others' actions. Moreover, the paired deficits reviewed in 4.2.2.1 support ST. Insofar as a deficit in having a particular kind of experience is paired with a deficit in interpreting others as having that experience, it seems likely that there is an involvement between the two processes. ST would predict this, but TT would not.

5.1.2 Comparing: Theoretical Considerations

Let me just call to mind some problems that each theory faces and which, I think, suggest, that each in fact requires elements from the other in order to work. I will start with TT.

5.1.2.1 Why TT needs ST

There are fairly good arguments to the effect that theory theory cannot work without incorporating elements of simulation theory, i.e. our own habits and dispositions. I will mention three of them.

1. Belief formation on the basis of *perception:* It seems reasonable that in figuring out what beliefs other people form about a situation on the basis of perceptions, we have to incorporate our own perceptions. I can't imagine how we could get around using this kind of use of simulation.

2. *The Frame Problem*: There is a lot of sensory information and also a complex web of beliefs and desires. How do we pick out the ones that are relevant in a specific case? Our own intuitions are a better guide than any formal technique as yet worked out by AI.[49]

3. *Rationality:* It is argued by some simulation theorists (in particular Jane Heal[50]) that a formal theory of rationality could not replace our own intuitions about what is rational. This can be significant for folk psychology in numerous ways. One point is that a theory could not capture the normativity that is inherent in our conception of rationality. We do not always know what follows from our beliefs; we sometimes have inconsistent beliefs; we have to make an effort to be consistent and are sometimes corrected, we exchange reasons to defend our inferences. Our judgments of rationality are therefore underdetermined by any formal account. So, insofar as we expect others to behave rationally, we have to rely on our own intuitions about what is rational. A second point is that the prediction of rational behavior in any given case concerns not just principles of rationality, but *the application or interpretation of these principles* in novel cases[51]. So what we would need is a theory that explains people's interpretation of the theory of rationality; and this theory would in turn have to be interpreted as well, etc. To avoid the regress, we have to use our own intuition about what is rational. A weaker form of the same argument would be that we always have novel constellations of beliefs and desires, so if we are to use principles to decide what follows from them the danger is that we need new principles for every novel case, and such principles are no principles at all.

5.1.2.2 Why ST needs TT

The obvious point, which I have already had occasion to refer to, is that successful simulation depends upon similarity to others. This does not mean that I cannot simulate someone who is relevantly different from me, such as someone of whom I know that she has a false belief. But in order to do so I have to identify the difference, and this I patently cannot

[49] This argument is worked out by Jane Heal 1996

[50] This one, too. And I also discuss it in the section on Jane Heal in chapter 4.

[51] Obviously, this is a very Wittgensteinian figure; I am not sure to what extent Heal may have Wittgenstein in mind.

do simply by checking what I myself presently believe or how I myself am presently inclined to act. Rather, I have to make an assessment about whether the target person is relevantly dissimilar to me. Picking out some feature and deeming it relevant, such as the fact that the target person was absent when the chocolate bar was re-located, is simply a different procedure from simulating the other person. It may *involve* simulating them – if, for example, I imagine myself seeing the chocolate bar hidden, leaving the room, and then coming back with the desire for chocolate. But how do I decide how many and which of the other person's experiences to simulate rather than just simulating their present situation? ST does not have any response ready-to-hand.

TT, on the other hand, has an easy answer, since it has it that my present prediction of the person's behavior (e.g. where they will look for the chocolate bar) depends upon ascribing beliefs to *them* on the basis of what *they* have perceived. Since, according to TT, we do not start out from our own first-person case and make adjustments but, rather, start out from a thrid person stance, there is no problem at all with the other person's being relevantly different from me. It seems that ST cannot get around incorporating procedures for making adjustments for differences, and these adjustments require me to step outseide of my first-person perspective and employ general knowledge about how people form beliefs and what people desire.

5.1.2.3 Comparison of the relative virtues TT and ST

Aside from these structural arguments, there are also some more general critical remarks to be made about each. Let me start out by looking at TT. One common criticism, for example, is that TT seems to be intellectualist to a degree that defies common sense. This is just the point that there is something fishy about the highly theoretical construal of folk psychology advocated by the theory theorists. When we are giving reasons for people's behavior, it seems more natural to talk about the world than about their minds. We say he is putting on his raincoat because it is raining outside and not because he has the belief that it is raining. ST may also appear a bit implausible from a phenomenolgical or common-sense perspective (Do we really go around all the time putting ourselve in others' shoes and imagining how we would act, etc?), but most people seem to agree that it is at least appears phenomenologiaclly to be less far from reality than TT. So this should count as a minor point in ST's favor.

The obvious TT response is to assert that the inferential processes referring to mental states occur implicity. But then, of course one has to spell out what one means by implicit inferences and implicit postulations of theoretical entities. One may be either persuaded or not about the answers that one gets to this challenge – i.e. one may either accept modularity or the analogy to Chomskyan linguistics, or determine that the empiricists' analogy between development of theory of mind and scientific theory change is a useful analogy. Otherwise there is no sense to be made of TT's use of the term "theory", which would obviously be a disaster for TT. Needless to say, ST has the same burden if simulationists want to appeal to the idea of implicit simulations. But ST is in a much stronger position here, since there have been lots of empirical findings in the years subsequent to the emergence of ST that look a lot like implicit simulations – e.g. mirror neurons, simulationst theories of concepts, Jeannerod's simulationist theory of action representation. We will be looking at these in subsequent chapters. For now I only want to point out that ST can actually regard these findings as successful novel predictions based upon the idea of implicit simulation, whereas TT does not have anything comparable to point to. So, this point, which amounts to *empirical fruitfulness* or productivity, speaks strongly for ST.

Looking critically at ST, the most important flaw is that it is a rather diffuse program. If we think back on Goldman, Gordon and Heal, it is clear that their approaches differ significantly, indeed more so than the various versions of TT. Goldman, for example, thinks it necessary to include introspection within the simulation procedure so that one can identify a mental state to ascribe to the target person. Gordon, of course, wants to do without this. Heal is talking almost entirely about ascribing rationality to others. We will see later on that Goldman, among others, talks about mirror neurons in terms of simulation. Other theorists propose simulatinist theories of conceptual thought and of action representation and planning. On the one hand, this lack of specifcity has to be considered a point of criticism. On the other hand, one might also conclude that the concept ot simulation has been productive and should be left open for the time being, so that it can be further developed in concert with ongonig empirical research.

Looking at other general theory-virtues, we can establish that TT is obviously less parsimonious than ST, since it postulates knowledge of specifically psychological generalizations at the core of folk psychology, which ST thinks that simulation procedures obviate. Thus, TT loses out on *theoretical simplicity*. This point should be separated from a related *ontological* point. Ontologically, TT may also be less parsimonious if it is interpreted realistically. Let me explain what I mean by this. According to TT, we postulate mental states in folk psychology as causes of people's behavior. One could be instrumentalistic, like Dennett, and not think that these postulates carry any ontological import, but then one has to explain why explanations involving these postulates work. This puts pressure on one to move toward a realistic interpretation, according to which there really are mental states causing behavior. Hence, TT at least urges one toward adding mental states to one's ontology. At least Gordon's version of ST is more parsimonious. To the extent that Gordon is right, we are not referring to any objects at all when we talk about minds or mental states and events. First and foremost, we are talking about things in the world from the perspective of someone who is experiencing them in a certain way.

Turning to specific feature of mental concepts that the two theories should address, the problem of *mental causation* is of course closely related to the ontological point just discussed. TT has a natural explanation for how the issue of mental causation arises, namely because mental states concepts are differentiated via functional roles carved out by causal relations to perception and action. That is, they are defined in part by the causal effects they have upon behavior. Note that my concern at the moment is not with which theory solves the problem of explaining how mental causation could be possible, but with the empirical issue of showing how this feature of mental concepts arises in the first place. It would be nice if our theory of folk psychology also established a starting point from which the philosophical issue could be seen more clearly, but that is a separate issue.

As for ST, on the other hand, there is no obvious account of how the issue of mental causation arises. Goldman's way of differentiating mental states is to suggest that they have different qualia. So, although he is free to say that they cause behavior, he cannot explain this link since the link to behavior is no longer a defining feature. Gordon and Heal do not even want to talk about mental states at all if they can avoid it, so obviously they is not going to be in a great position to say much about mental causation. They would probably say that the issue is a philosophical construct and not really inherent in our everyday concept at all, and thus a pseudo-problem. Depending on your intuitions, you might consider this either quite reasonable or extremely unsatisfying.

This very same move that enables TT to account for the ssue of mental causation, on the other hand, makes it hard (but not impossible) to account for *privileged access*, since the functionalist view of mental concepts implies that we would have to self-ascribe mental states by considering what we have perceived and how we have acted. ST, on the other hand, gives a better account of privileged access, since it makes ascription to others dependent upon self-ascription. This is true in a weaker sense for Heal and Gordon, since they think at least that knowing how others will think or act depends upon knowing how one would think or act oneself, although they do not phrase this in terms of privilged access *to one's own mind*. It is true in a strong sense for Goldman, since self-ascription has nothing to do with postulating functional states but is a matter of introspecting what state one is in on the basis of qualia, which are given only to the experiencing subject in the first-person. Obviously, this gives Goldman's version of ST a clear advantage with respect to accounting for how qualia enter into our conception of mentality. TT has at least the difficulty that qualia do not directly figure in the definitions of mental states and play no role in differentiating among them, which seems counter-intuitive.

What about *intentionality*? I will say more about intentionality in chatper 7, where I embend TT and ST in broader theories of concepts. For now, just a few remarks. Empiricist versions of TT focus on the representational nature of mental states, and thus naturally incorporate intentionality. This is not necessarily true of nativist theories, but both empiricist and nativist versions fairly naturally incorporate accounts of intentionality because they endorse functionalist theories of mental concepts that define mental concepts by virtue of links to the world. In fact, they fit best with accounts of teleosemantic accounts (Dretske 1981, Millikan 1984) according to which the intentionality of a mental representation consisting in a causal link to things, processes or events in the world and to its having the function of standing in such a causal relation. This latter, functionalist, addendum makes it possible to make sense of the possibility of misrepresentation, which everyone nowadays agrees is a necessary condition for being a representation.

Goldman's introspectionist approach to mental concepts gives him problems here for just the same reasons as in the case of mental causation: mental concepts, for him, are defined by qualia rather than by any link to the world. He can build in such a link, but it is not automatically there as it is for TT. Although Gordon avoids talk of mental concepts, or spells them out in terms of states of affairs in the world (i.e. replace the belief that p with the assertion p), this link to the world does give him a foothold in the discussion of intentionality. Gordonesque explanations of behavior imply intentionality of people's decision-making and action planning, since they are formulated in terms of the intentional content of those mental processes. For example, Gordon would urge that the natural explanation is "Jones took his umbrella because it was raining" rather than "Jones took his umbrella because he believed it was raining." Thus, the propositional, or intentional, content ("that it was raining") is playing the decisive causal role in producing the umbrella-taking; it is just that the propositional attitude is left out. So, implicitly, Gordon's theory naturally incorporates intentionality. In fact, as we will see in chapter 7, Gordon and Goldman (as well as Heal) are welcome to the same teleosemantic account of intentionality as TT, but in order to see why we will have to wait until chapter 7.

5.1.2.4 Summing Up

So, as we have seen, differentiating empirically between the two theories is hard or impossible. ST has the advantage of having made novel predictions that have been successful (in particular mirror neurons and other mirroring phenomena, which will be discussed later on). Theoretically, there is no clear winner, but both sides have some stroing points and some weaknesses. One may therefore want to throw both out and start over or else start thinking about how to combine insights from both theories. We will take a look now (5.2) at some approaches that endorse the first option, and then turn to the second option (5.3).

5.2 Condemning theories of folk psychology (maybe there is no such thing as folk psychology)

Earlier in this chapter, we started out from the viewpoint that neither ST nor TT alone suffices to account for social cognition in general and focused on various options for combining them to form a hybrid theory. Another way of moving forward from that viewpoint is to deny that TT and ST are useful conceptions at all and to assert that they should be scrapped rather than combined. This position has in fact gathered support over the past ten years, in particular among philosophers who are interested in the phenomenological tradition as well as current empirical psychology and neuroscience. Shaun Gallagher and Dan Zahavi are two particularly prominent examples.

5.2.1 Gallagher's criticism

Gallagher has been an influential proponent of the view that this whole folk psychology business is a load of bunk. Although he has criticized ST a bit more prominently and explicitly, this does not mean that he finds TT better. Far from it, his view seems to be that TT is some kind of intellectual joke that was mistaken for a theory by some really humorless bozos, whereas ST is a legitimate position worth criticizing. Although he does not say this explicitly anywhere, it is pretty clear that his diagnosis is that ST suffers from the fact that it arose as an alternative to TT and as a result shares certain presuppositions with TT that are just plain silly and should be thrown out the window rather than re-packaged in an alternative theory. What presuppositions are we talking about?

Silly Presupposition #1: Gallagher refers to this as the mentalistic supposition, and formulates it thusly:

> 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 (2005, 209).

This presupposition implies, according to Gallagher, that what we are up in social cognition crucially involves mental concepts, which we use to understand what is going on inside other people's minds. He admits a distinction between explicit recognition of another person's mental states, which is "clearly conceptual", and implicit recognition, which is "informed by such conceptual knowledge." That is fair enough, since this implicit version of recognition could be interpreted in a pretty deflationary way that most participants in the debate would agree with[52], but Gallagher instantaneously abandons this reasonable distinction and speaks of ascription in pretty robust terms:

> To discover a belief as a intentional state even in myself requires that I take up a second-order reflective stance and recognize that my cognitive action can be classified as a belief (2005, 209).

And he goes on to conclude that we do not usually do seem to do this. It does not seem to be a good description of our phenomenological experience. Right, but maybe we do it implicitly. A proponent of ST or TT could say that the explanation they are giving of social competence in terms of belief ascription is a *functional explanation* that is not intended to be true to our phenomenological experience. What belief ascription, for example, means in this sense is simply to form certain expectations about others' behavior on the basis of what the person appears to have seen, heard, said, done or whatever.

But Gallagher is right, I think, to question the emphasis on cases where the concepts of belief and desire would plausibly be needed (such as false belief tests). It is a fair point to say that these cases might be the exception, and to focus too much on them could be misleading. But it is not really fair to say that everyone on both sides has fallen prey to this error. What he says in fact sounds a lot like what Gordon emphasizes 'round the clock. Which means that he opens himself up to the same kinds of criticism that Gordon faces. At some point, mental

[52] But: Can the reader guess which prominent advocate of ST would not agree? See chapter 3.2, or else just read a couple paragraphs more here and you will get it…

concepts do have to be introduced in order to account for people's ability to correctly anticipate the behavior of people who differ from them, in particular to notice and account for relevant differences in beliefs and desires.

And aside from the version of ST espoused by Gordon, who, as we know, wants to do without mental concepts for as long as he can get away with it, there are also theoretical options that include mental concepts, but more in the vein of implicit recognition. Gallagher gives no reason for ignoring the possibility of an account of folk psychology that characterizes ascription as involving concepts, but in an implicit sense that saves the phenomenology a bit better than the caricature-version Gallagher mentions. Embodied concepts, as we will see in later chapters, present such an option. The funny thing is, Gallagher himself proposes a version of this, as we shall see below.

Silly Presupposition #2: Gallagher goes on and on an on about how nutty it is to think that what we are doing in everyday situation is best characterized as explanation and/or prediction. The disinterested, scientific flavor of these terms does not fit our everyday experience of social interaction. According to Gallagher:

> …what phenomenology tells us is that explanation and prediction are specialized and relatively rare modes of understanding others, and that something like evaluative understanding about what someone means or about how I should respond in any particular situation best characterize most of our interactions (Gallagher 2005, 212).

As for the key term here, "evaluative understanding", he does not give us too much help. As far as I can tell, it is the social-cognition-version of "pragmatic interaction" – a term that phenomenologists apparently use to describe our primary and usual mode of relating to the world around us[53]. What is characteristic of this mode is that we usually encounter things (and also people) in the course of activities that come naturally to us and about which we do not have to think too much. We only start predicting or explaining when something or other goes awry.

This is all well and good – and we will come back to it when we discuss Gallagher's positive proposal below – but the obvious comeback for anyone defending ST or TT is that psychologists and cognitive scientists are looking for explanations of how we manage to interact with people (and things, of course, but that is a separate issue) as we do, not for descriptions of how it seems to us. The explanations can refer to unconscious physiological processes, or they can be functional explanations that remain neutral with respect to the neural

[53] Actually, he says "being in the world".

substrate, but there is at any rate no reason why they should have to refer to the conscious processes that phenomenologists would be phenomenologically interesting or accessible.

But he is right about one thing. If we only mean the terms "explanation" and "prediction" (the same goes for "theory", "simulation", "ascription" and surely a bunch of others) in a more or less metaphoric sense, then we are not entitled to depend to heavily on the analogy to real prediction and explanation in order to make sense of what we are talking about. The upshot is that we should aim either for physiological explanations or for more specific functional explanations that, and should avoid terms that imply conscious experiences.

Silly Presupposition #3: Another presupposition of TT and ST that Gallagher wants to expose is that social cognition is primarily third-person observation rather than second-person interaction. In other words, interactive engagement with others, as in cooperation, joint attention, emotional contagion, offers us ways of gaining understanding of them that we would not otherwise have, i.e. in detached observation. This is a good point. Theoretically, it fits well with the widespread view, which we will discuss later on, that an important difference between humans and other primates is that we are more interested in engaging with others socially (e.g. cooperation, shared attention, etc). By ignoring these effects, we could be neglecting forms of social cognition that are characteristically human. That would be foolish. We will be discussing these issues at greater length in connection with Tomasello's[54] views on cognitive development. For now, I will limit myself to a few remarks that bear on Gallagher directly.

He himself suggests that false-belief tests are flawed in their focus on 3.person understanding. The child being tested typically watches some other people (or puppets) with which she does not interact, and then answers questions posed by an experimenter. We could make these scenarios more interactive. I think he is absolutely right that this is a shortcoming, and there is no reason why false belief tasks should not be constructed that investigate whether children fare better in cooperative contexts.

Gallagher's own positive account, which he calls "interaction theory", takes its departure from ways of relating to others that are generally regarded (at least according to Gallagher) as "precursors" to theory of mind skills. Gallagher's move is to assert that these "precursors" in fact constitute the bulk of our mature adult interactive techniques, and that the processes characterized by TT and ST only come into play in special cases. So what kinds of

[54] See Tomasello et al. 2005

"precursors" are we talking about here? Gallagher's account is based on a structure proposed by Trevarthen (1979), who distinguished between primary and secondary intersubjectivity.

Primary intersubjectivity is characterized by simple mechanisms that enable children to engage with other people in ways that differ from their interaction with inanimate objects. Neonate imitation is one famous example. Meltzoff and Moore (1977, 1994) have shown that neonates have a tendency to imitate specific facial expressions more or less immediately after birth. The million-dollar question, naturally, is how they match motor representations with perceptual representations. Gallagher, in any case, does not think that inferences have to be drawn or models compared. He ascribes neonate imitation to the workings of the body schema, understood here as an "innate system designed for motor control", which is constituted by "a set of pragmatic (action-oriented) capabilities embodied in the developing nervous system."(Gallagher 2005, 226) This is not overly lucid; what I believe he has in mind is that there is a common code that represents one's own motor schemas as well as perceived movements of others. There is an innate association of one's own motor schemas with certin perceived objects and movement patterns. Something that looks like a tongue and/or moves in a certain way simply activates an infant's tongue. In the infant, this is already intact, and the match is improved gradually with proprioception and visual perception of others and of one's own body. But there is no need for inferences involving abstract representation of actions or mental states, or a sophisticated self/other distinction.

Gallagher adds to the picture two of Baron-Cohen's modules[55]: intentionality-detector (ID), which enables children to distinguish intentional action of an agent from non-intentional movement, and the so-called eye direction detector (EDD), which enables children to recognize where another person is looking. With regard to the latter, he departs from Baron-Cohen on one point insofar as he denies that complex inferences have to be drawn about what the other person sees, what they know or do not know on the basis of what they see, etc. Infants presumably can get pretty far using simple behavioral rules concerning how others act on the basis of where they are looking. Their own experience of looking in a given direction and being aware of what is there, he notes, may provide the basis for this understanding of others' experience of seeing. But we need not explain this in terms of an analogy from my mental states to yours. It could work something like this: The ID picks out intentional agents; the EDD picks out an object they are looking at; automatically an expectation is formed that those agents will account for that object in their actions.

[55] See above, 3.3.2.

Another issue here is whether an inference has to be drawn from looking at an object to seeing that object. Baron-Cohen (1995, 43) thinks so, since it is possible to be looking toward an object without seeing it, either because one's eyes are closed or because one is distracted, etc. So an infant would have to gather additional information (are the other person's eyes open?) and draw upon various personal experiences (having one's eyes open versus closed) and draw an inference. Gallagher does not agree, and his counterproposal is sound. It is more plausible, he notes, that the default mode of the EDD is to treat someone as seeing something when they are looking toward it, and that the distinction is gradually introduced as more experiences are gathered. Hence, treating someone as seeing something would not require an inference from their looking at it.

In other words, it is possible to describe these cases without phrases like "simulation", "analogical inference from me to you", "using oneself as a model of the other" and so forth. If we accept for a moment[56] that some of the same neural networks are used in planning and executing actions, on the hand, and in observing actions on the other hand, then improvements in one's own understanding of one's agency could automatically become available for social cognition, but without any explicit inference being drawn. The invocation of notions like simulation or analogical inference is therefore an optional gloss, and Gallagher is justified in wanting to be skeptical about such loaded language that suggests conscious processes.

Finally, he notes phenomena of "emotional attunement". Infants "vocalize and gesture in a way that seems 'tuned' to the vocalizations and gestures of the other person" (Gopnik and Meltzoff 1997, 131). Moreover, at 5-7 months, they detect correspondences between visual and auditory expressions of emotions (Meltzoff 1997, 131). One might add that infants engage in so-called "protoconversations" pretty soon after birth. Trevarthen (1979) describes these as social interactions in which the infant and the parent – through looks, touches and vocalizations – express and share basic emotions. So there is a tendency for their emotional states to correspond with each other. Moreover, the interactions have a turn-taking structure.

What characterizes *secondary intersubjectivity* is that the interaction between the infant and the other person is extended to include an event, an object or an activity that they attend to together. Paradigmatic for this type of situation is what is called "shared attention", which develops around 9-14 months.[57] Typically, the infant will monitor the gaze of the other person, checking to verify that they are continuing to look at the same thing. This same interest in shared attention is also exhibited in the phenomenon of pointing, which becomes increasingly prevalent around 11-14 months (Tomasello 2008). Just as an example:

[56] This will be dealt with extensively in later chapters, e.g. chapter 6 on mirror neurons.

[57] See Baron-Cohen (1995), who postulates a "shared attention mechanism" (SAM).

Liszkowski et al. (2004) did a study in which declarative pointing gestures were provoked from 12 month olds (novel objects suddenly appeared at some distance from them). They pointed at the objects and looked to adults to see if the adults would look at the object. If the adults did not look, they became agitated and continued to point; if the adult looked at the object but then not back at them, or did not express interest, they continued to point; if the adult looked in the right direction and then back at them to express interest, but could not have seen the object because it was blocked from their line of sight by an occluder, the infants continued to point. Only when the adult looked at the object (and, if necessary, moved in order to be able to see it, and then looked back with interest, did the children appear satisfied and stop pointing. This pattern suggests a sophisticated understanding of others' emotional reactions to situations and of their epistemic grounds for reacting as they do. And yet these children will not be able to pass false belief tests for a couple of years. So it seems reasonable at least to look for an account of the social cognitive sophistication expressed here that avoids mental concepts.

Of course, Gallagher acknowledges that beliefs and desires have to come into the picture at some point. But, as I mentioned above, he thinks that we can characterize them in a thinner, more deflationary way. Specifically, he suggests that we do not need to think of beliefs as "all-or-nothing mental representations" (214), or to postulate "idealized and abstract representations standing behind these behaviors in order to grasp the disposition that is overtly constituted and expressed in the contextualized behavior" (215).

Instead, he thinks we ought to regard beliefs as a "more or less complete set of dispositions to act and to experience in certain ways" (214). The "and experience" is an interesting tip: Gallagher asserts that one can have dispositions to have particular phenomenal experiences, e.g. to "feel upset". It is unclear whether Gallagher thinks that such phenomenal events constitute part of the set of occurrences to which particular beliefs dispose us, and which therefore also partly define beliefs, or whether he just has desires in mind, or what. But it seems to me more plausible to include phenomenal experiences if we are going to say that beliefs are defined by what they dispose us to. Just take a simple example: Judy has the belief that Elvis is dead. This disposes her to be surprised, delighted, scared or whatever if she sees him. As far as I can tell, this disposition is just as much a part of the concept of a belief as the disposition to, say, search under the couch for my socks because I have the belief that they are under the couch. Of course, this does raise the issue of what the criteria are for establishing that a disposition has been actualized. Phenomenal experiences are only accessible from a first-person perspective, so we will have to give some account of how the criteria are nevertheless the same – or at least pretty close to the same – for just about everyone. One option would be to put together an account of how people learn to correlate publicly observable expressions of phenomenal experiences with their own first-person experience. Another option would be to back off just a bit and say that beliefs, conceived as dispositions, are actualized not by phenomenal experiences themselves but by the publicly observable expressions of these phenomenal experiences. But this is not the place to pursue this issue any further.

Gallagher's gloss on all this is that children's interaction with humans differs from their interaction with inanimate objects on the basis of mechanisms that do not require a conceptual understanding of others' mental states. What they are doing is not mind-reading but body-reading (Gallagher 2005, 230). Moreover, he thinks it likely that these same mechanisms – which others may regard as mere precursors to mature theory of mind skills – constitute the social-cognitive toolbox that we continue to use as adults, and then processes like simulation and theorizing come in only in rare situations in which the normal techniques do not work.

Summing up on Gallagher, I would venture the claim that his theory is not so terribly different from some versions of TT and ST – in particular from Perner (TT) and Gordon (ST). Recall that Perner thinks that older children and adults remain "sutuation theorists" at heart, only using metarepresenstational skills to represent others' mental states *as mental states* in rare cases. Gordon also thinks that most of folk psychology is a matter of correclty anticipating how people act in particular situations without having to think about the mental states that intervene in causing their actions. The accounts that these two give differ from each other and from Gallagher with respect to what elements they introduce to account for special cases where the target person's behavior is opaque (such as false belief tasks), but Gallagher exaggerates these differences by giving the impression that TT and ST were committed to the claim that either explicit theorizing or explicit simulating were the predominant method. And, as we have seen, some (especially Perner and Gordon) think that in the majority of cases we simply anticipate correctly what the other person will do. Talk of theory or simulation is intended to explain how this can be, and appeals to implicit and/or unconscious processes. We will say more later in this chapter and also in subsequent chapters about the notion of implicit application of theory or simulation. Suffice it to say, for the moment, that one can as a theory theorist or simulation theorist agree with Gallagher that what we usually experience is that we just know how others will act or understand why they acted as they did without making any effort. As a theory theorist or simulation theorist, one can agree with everything

Gallagher says and still insist that some additional explanation of the implicit and/or unconscious underlying these acts of understanding would still be desirable.

5.2.2 Dan Zahavi

Zahavi's critique can be seen as complementary to Gallagher's. Like Gallagher, he agrees with ST that TT is just plain nuts but thinks that ST does not go far enough. He, too, questions the notion that explanation and prediction are at the core of our everyday social interactions, and rejects certain fundamental assumptions shared by ST and TT. In particular, he thinks that both theories take their departure from the view, "defended by Cartesians and behaviorists alike" (518), that "our encounter with others is first and foremost an encounter with bodily and behavioral exteriorities devoid of any psychological properties"(518); i.e., that others' minds are not directly experienced. According to this view, our access to other minds must proceed by some kind of inference, i.e. either by theorizing about the unobservable causes of their behavior or by drawing an analogy from our experience of our own minds. The basic assumption, Zahavi argues, underlies the entire discussion of TT and ST.

In his critique of this view, Zahavi draws upon certain distinctions and observations made by Max Scheler. This sort of thing (appealing to phenomenologists) is by no means as unusual as it was twenty years ago or so in analytic philosophy. As I mentioned in the introduction to this chapter, Gallagher and Zahavi are both examples of philosophers who simultaneously stand in the phenomenological tradition and work with empirical psychologists and neuroscientists. Although a thorough discussion of the phenomenological background that these philosophers draw upon would take me too far afield, Zahavi presents an opportunity to look briefly at the way in which phenomenology can be brought to bear upon the issues at stake here, so I will sketch the main points that he borrows from Scheler.

In particular, he discusses a sort of classification of related but distinct states in which we have direct experience of others' emotions and an emotional response to them. The most important of these is a basic state, which Zahavi dubs *empathy*[58]. The term "empathy" designates a direct experiential understanding of others' minds; a "basic, irreducible, form of intentionality that is directed towards the experience of others." In Scheler's view, this kind of access to others' emotions is a matter of perception, and he refers to his theory as "a

[58] Zahavi notes that Scheler does not stick to a single term in referring to this more basic state, but switches among various terms, such as *Nachfüllen, Nachleben, Nacherleben, Verstehen* and *Fremdwahrnehmung*. Since my interest is in what Zahavi does with this phenomenology and not in interpreting Scheler, I will not address the different nuances of these terms. It is only perhaps worth pointing out that the German word *Einfühlung* – the German word one would most readily expect – does not appear on the list. Zahavi explains that this is because Scheler wanted to distance himself from Lipps' projective theory of empathy (Zahavi, 516).

perceptual theory of others' minds" (Scheler, 1954, 220). I will come back to this idea of perceiveing others' mental states below (in about 1-2 pages and then again in chapter 7), since it presents an interesting parallel with some other theorists who do consider themselves more or less in line with ST, such as Elisabeth Pacherie, Marc Jeannerod and Joelle Proust.

Anyway, Zahavi and Scheler are on the same page with ST so far. Empathy, according to Scheler, is distinct from "intellectually judging that someone else is undergoing a certain experience" (Zahavi 517).). Obviously, this sounds a bit like ST so far, insofar as it is distinguished from an "intellectual access" to others' minds, which one is tempted to associate with TT. Unfortunately, Zahavi does not tell us exactly what the rejection of an intellectual judgment is meant to reject (psychological laws? concepts?).

But there is an important difference between this conception and ST. Perceiving or experiencing others' minds, for Scheler, does not have to proceed via an analogy to one's own mind. Scheler is criticizing the argument by analogy espoused by Lipps and Dilthey, which Zahavi deems the "grandmother of the kind of simulation theory espoused by Goldman" (518).[59] According to Scheler (at least as Zahavi reads him) the argument from analogy underestimates the difficulties of knowing one's own mind and overestimates the difficulties of knowing others' minds. Importantly, experiential understanding of someone else's emotion does not necessarily involve being in the same state as the other person. You may wind up experiencing the same emotion as the other person, but – contra Goldman – this is not a prerequisite for having a direct experiential understanding of it. An example from Scheler intended to illustrate this point is that of a sadist, who is not ignorant of his victims' pain and does not have merely intellectual knowledge of it, but experiences it in an emotionally engaged way and indeed empathically enjoys it (Scheler 1954, 14). Hence, simulation cannot explain our ability to understand others' emotions.

I do not think that the example is to terribly persuasive. A simulationist has at least two possible responses. She could go far a psychological explanation and maintain that the sadist's pleasure somehow includes, sublimates or developmentally presupposes the experience of pain that he inflicts on others. Or she could opt for a neuroscientific or biological route and wager that the sadist does in fact have an experience (or a partial experience) of pain that can be demonstrated in neural, muscular or hormonal activity.

But I would not deny that the departure from ST on this point does seem plausible for some other examples. Zahavi points out that when a furious assailant lunges at me, I can

[59] Analogy in the sense of projecting our knowledge of our own minds onto others. Of course, not all simulationists, and certainly not all theory theorists accept that this kind of analogy plays a significant role in folk psychology.

understand that he is furious without getting furious myself. I may for example become terrified (Zahavi, 517). A simulationst may respond that in some cases like this one, I can respond to someone's behavior appropriately without understanding it. Evolution has fitted me with some automatic stereotypical reactions to stereotypical situations that spare me the trouble of understanding others. We could go back and forth on this until we are blue in the face (and, in fact, if someone pays me to do so, I would be happy to), but there is no need to expand the point here, since this issue will be central to the discussion in chapters to come. For the moment, let it stand that Zahavi and his buddy Scheler assert, like ST, a direct experiential understanding of others' minds, and yet have valid grounds for doubting that this kind of understanding must involve being in the same state as the other person.

Sympathy is a similar state insofar as it involves a direct experiential understanding of someone else's suffering, but it is different in that it also involves a more specific kind of emotional response, which Zahavi characterizes as "compassion or concern" (Zahavi, 516). I am not so sure that "compassion" is all that helpful as a definition of "sympathy", but "concern" is certainly better. Another similar state (or, let's call it a phenomeneon) is emotional contagion. You go into a bar and everyone is jolly and you become jolly but do not notice, at least in any conscious sense, that anyone else was jolly. You may not even realize consciously that you yourself are jolly. This phenomenon differs from empathy and sympathy in that it does not involve understanding anyone else's emotional state or being aware at all that one's state is related to anyone else. But, and here's the kicker, it does involve being in the same state as someone else. One may see this as yet another reason to resist thinking – like Goldman – that understanding others' minds involves being in the same state as them, or mirroring them (Zahavi does not say this). It is not a very good basis for criticism, though, since Goldman could replay that mirroring is a necessary but not a sufficient condition for simulation[60].

5.2.3 A few "thinking outside the box" approaches (Ratcliffe Bruner, Kusch, Knobe)

A number of philosophers (and also some psychologists) have recently joined the ranks of the critics who think that the very terms of the folk psychology debate are misconceived and should be re-thought altogether. Many such perspectives are included in a volume edited by Daniel Hutto and Matthew Ratcliffe (2008). In this section, I will briefly look at a few of these proposals/cirtiques.

[60] In fact he does say something in this vein (see chapter 6).

Matthew Ratcliffe (2007) takes much the same line as Gallagher and Zahavi, arguing that the very notion of a folk psychology involving the ascription of beliefs and desires is a mere philosophical fiction that does not reflect reality and should be cast to the flames. He adds a bit to the argument for this view by having actually done some empirical studies designed to test whether the elements of folk psychology match what normal people take themelves to be doing in everyday life. So, for example, he asked undergraduates who had not yet heard of folk psychology to respond to the following question:

> What is central to your understanding of others? To put it another way, understanding or interacting with another person is very different from understanding or interacting with a rock. What does that difference consist of? Please state your intuitive or commonsense view rather than stating philosophical positions or engaging in philosophical argument.Write up to half a side of A4 and return it to me at next weeks's tutorial (Ratcliffe 2007, 47).

In 25 responses in the first study, the term 'belief' appeared only twice, and 'desire' and 'prediction' only once each, while 'prediction' was never mentioned. The next time around, the results were similar. Ratcliffe acknowledges, however, that some responses sounded a bit like ST – e.g. "I assume that other people are essentially the same as me", "That they think in a similar way", etc. (2007, 48).

This is definitely interesting and challenging to anyone involved in the folk pyschology debate, but the results are more limited than Ratcliffe seems to think. They suggest strongly that we do not predominately use simulation or theory in an explicit straightforward sense. But they do not unmask the entire folk psychology debate as nonsense, since theory theorists and simulation theorists still want to understand the processes that enable us to understand others in a way that seems direct, intuitive, etc. Patently, there are *some* cognitive processes going on. The theory theorist or simulation theorist will simply insist that TT or ST give the best account of those processes, even if we have no conscious awareness of or control over the processes in question. In summary, Ratcliffe puts pressure on TT and ST to spell out what is meant by the notion of implicit theory or implicit simulation, since his results suggest that there is not much explicit theory or simulation going on.

Jerome Bruner (1990) is similar to Gallagher, Zahavi and Ratcliffe in that he emphasizes the importance of contextual and narrative information in making sense of other people's behavior. Coming from a psychological background, Bruner winds up with some different, equally interesting points. For Bruner, one especially central point is that we can understand people's behavior in terms of social roles. This is different from thinking about an immediate situation, and also different from thinking about what is going on in people's heads. Instead, we think about people's roles, about the kinds of behavior they produce on the basis of their jobs, their social positions, their relationships to other people in a given situation, etc. So, for example, we would explain why Hans gets up at 4:00 in the morning by noting that he is a baker and bakers have to start baking at ridiculous hours. We do not have to think about his mental states in order to given such an explanation. On the other hand, one could analyze this explanation by saying that it appeals to beliefs and desires implicitly (e.g. He desires to bake bread or to retain his job, etc.), but Bruner's point would be that this analysis introduces more complexity than is probably necessary. We probably use a lot of this kind of social knowledge in a fairly automatic way, without in any interesting sense having to think about beliefs and desires or simulating the perspective of a baker.

Martin Kusch (2006, 2008) has critized the folk psychology discussion for slightly different reasons. His target is the notion that mental states such as beliefs and desires are to be conceived in what he calls an individualistic fashion. By this he means the view that they are isolable states, objects or processes in people's heads that cause their behavior. Kusch thinks that they are better conceived as social constructs. He thinks ST is right that our similarity ot others helps to make their behavior intelligible, but denies that this must involve special processes whereby we simulate them. Instead, he would appeal to narratives and roles that we all know in common.

The main point that he raises, which I find quite reasonable, is that, as he puts it, "folk psychology has a self-referential component" (2006, 326). What he means by this is that the truth or falsity of mental state ascriptions is dependent upon criteria that we decide upon as a community. At the core of this dependency upon the community, for Kusch, is the normative/ rational component of folk psychological ascriptions. As we have discussed elsewhere, ascribing someone beliefs and desires in light of their actions, or predicting actions in light of ascriptions, involves assuming that she is rational. If we believe her to be thirsty and she refuses a drink, for example, we will conclude either that we were mistaken to think she was thirsty or that she is behaving irrationally[61], since one does not qualify as being thirty if one is not inclined to accept a drink. That is just part of the concept of thirst. So the truth of the ascriptions depends on the grammar of concepts established by a linguistic community.

Hence, according to Kusch, it is misleading to say that mental concepts refer to things in people's heads. Rather, their reference is a complex affair that crucially involves individuals' actions as well as the communities in which these actions are interpreted. I think

[61] Obviously, we will first look around for other ad hoc explanations (e.g. she is being paid by NASA to go without fluid for as long as possible); only if we find no such explanation do we have to send her to the madhouse.

this conclusion is quite reasonable, and I will take it up again in the final chapter. But Kusch, like the other theorists in this section (5.2), does not answer to the demand made by TT and ST to account for the (probably largely implicit) processes that underlie our ability to make mental state ascriptions. There is no reason why TT and/or could not endorse Kusch's conclusion about the self-referentiality of mental states or social constructedness of mental concepts.

To conclude this section, I would like to mention one last unconvential perspective upon the discussion, namely that introduced by Joshua Knobe (2006, 2008). Knobe has established a whole series of studies revealing that moral evaluation plays a prominent role in folk psychological concepts. The original example used in the first experiments involved a vignette that people were confronted with in which an executive was told that a particular program would bring profits but would also harm the environment. The executive decides to go ahead with the program, and indeed they make profits and harm the environment. The test subject are asked whether the executive intentionally harmed the environment, and they overwhelmingly say 'yes'. But if the vignette is altered ever so slightly, so that the executive is told that the program will help the environment, and indeed it does, the overwhelming majority respond that the exectuive did *not* intentionally help the environment.

Since the two vignettes are structurally the same, it is surprising that people regard the deleterious action as intentional and the helpful one as non-intentional. Knobe argeus that this reveals that the very concept of intentional action crucially invovles not only description but moral evaluation. We use the concept in order to make moral judgments, not just to make predictions or give explanations. This moral dimension, according to Knobe, has been totally left out of the folk psychology discussion and should be taken seriously. Moreover, it reveals a limitation of the analogy to science since science does not have the aim of making moral judgments about what it investigates.

Knobe's findings (which have been replicated and followed up in numerous contexts in various cultures) clearly reveal an important element of folk psychological concepts that, as I argue, should be taken upon within the discussion. At the very least, it should be acknowledged as a component of folk psychology. Beyond this, obviously, it would be interesting to explain why there is this particular assymetry that Knobe mentions, and what evolutionary, psychological, social, or reasons there are for it. But so far no one has come up with such an explanation.

5.3 Combining: Theoretical Considerations

We have seen that TT and ST both have their problems, and indeed that each seems to need the other in one form or another. We have looked at some approaches that want to depart radically from the framework of a debate between two rival theories. It also seems reasonable to consider searching for a hybrid theory, and this has in fact been the trend in the past 10 years or so. I would therefore like to present some theoretical ideas about hybrid theories, and then turn to some hybrid accounts.

One occasionally hears people express the view that one or the other theory might be in some sense primary. I start out (5.3.1) by considering what this could be taken to mean. Secondly (5.3.2), one also occasionally hears that they might both be implicit. What could this mean? Thirdly (5.3.3), I will consider whether the analogy to science, which is invited by the names of both theory theory and simulation theory, can help to shed light on the ways in which theory and simulation can productively be combined. Then (5.4), I will discuss some hybrid positions that have been proposed.

5.3.1 The primacy of TT or ST

Here are a few ways of interpreting the claim that either TT or ST is primary with respect to the other.

Sense #1 - *A developmental thesis*: i.e. that one strategy arises earlier than the other. If ST is primary, the developmental sequence could run like this: children begin by expecting others to know and want the same things they want, and then gradually introduce corrections, such as recalling that Mommy likes broccoli more than cookies. This correction may not need to involve a very sophisticated understanding of mind, since it could be learned by plain-old association. More mentally sophisticated corrections may be required for things like false-belief tasks. If TT is primary, it could be that children start out by making observations and noticing regularities, and subsequently learn to correlate their own first-person experiences with outwardly observable criteria, and from then on can use their own first-person experiences as a basis for simulating others. So, for example, they notice that people who yawn tend to go to bed, and postulate some inner state (call it S) that causes yawning as well as going-to-bed-behavior. So, they expect yawning people to go to bed. Subsequently, they correlate their own yawning and going-to-bed-behavior with feeling sleepy. Feeling sleepy play the same role as S, so they are no in a position to conclude that yawning people are sleepy and form predictions about sleepy people's behavior on the bassis of their own experience of sleepiness.

Sense #2 - *One strategy could be used in acquiring the other*: This is related to the first option. Theory could be used to acquire or improve simulation skills in several senses. For example, by making observations and learning regularites, one makes sense of people's actions and can then engage in those same actions oneself, whereby one then has first-person experiences that can be used as a basis for simulation. Alternatively, children may use simulations to learn theoretical knowledge, or knowledge about people's roles and personalities, typical narratives, conventions, etc., and then employ this knowledge without in any interesting sense performing a simulation at the moment when they apply this knowledge.

Sense #3 - *Default strategies*: If simulation is the default strategy, we could *as adults* make the default assumption that others behave as we would in their situation, and then correct for differences by introducing theoretical elements as necessary. Alternatively, if TT is the default strategy, we could use the theoretical apparatus in normal cases, but revert to simulation when the theory fails. This model works best, I think, if we think of the theory part as being implicit, automatic and unconscious, and the simulation part as being a deliberate act of imagination. So, for example, when some new acquaintance fails to return my calls, it may surprise me at first, since I have generally observed that people call back when you ask them to. Then I would begin to attempt actively to interpret his behavior. One thing I might then do is put myself in his shoes and see what might lead me to decide not to call back (e.g. if I thought the other person wanted money, sex, etc.).

Sense #4 - *The endpoint of folk psychological explanation*: What I mean by this is that we could apply one strategy as a sort of heuristic or auxiliary to the other, and regard the latter as the one that really satisfies are search for an explanation. So, if theory is the endpoint, we could fiddle around with simulations in order to to come up with hypotheses, but regard our search for an explnaation complete when we come up with a theoretical formulation that fits the data – e.g. "he is angry and angry people presented with punching bags tend to punch punching bags, therefore he punched the punching bag". If simulation is the endpopint, it could be that when we do introduce theoretical elements into our considerations in order to make others' behavior intelligible, we do so in order to be able simulate their perspective imaginatively. Simulation, then, would be necessary in order to make use of theoretical elements. Let's take a false belief scenario for example. If I realize that you have different beliefs from me, namely a false belief, then I still need to imagine myself in your situation with that false belief in order to predict what course of action is reasonable on the basis of the false belief.

Sense #5 - *One strategy implicitly underlying the other*: It could be that one account (ST or TT) gives an elegant description of how somebody understands somebody's else's behavior in some everyday context, but that the other account gives a better explanation of the underlying mechanisms, for example because this account can be applied more broadly to a greater range of cases. I will say more about this in section 5.3.2.1, where I discuss models like this proposed by Ravenscroft (2008) and Nichols and Stich (1992). But before coming to that, I want to offer a few ideas about the use of the term "implicit" in this context.

5.3.2 In what sense could simulation and theory be used implicitly in folk psychology?

It is important to clarify one point – namely, that we can think of simulation and theory as tools that are used *implicitly* in understanding others. This strikes me as an important point since other people's behavior is often predictable without our having to think about it – that is without having to go through complex processes of ascribing unobservable mental states and applying generalizations about behavior and then deriving predictions, and also without actively imagining ourselves in their situation, perhaps coordinating various imaginative representations, one of a situation where we are angry and another of the other person's situation. If we are looking for a description of how we really go about understanding people in everyday life, then we do stuff like that only when we have to. If the person's situation is like one we have often experienced, then we just know how they will act. Or if they are relevantly different and we know in what way, then we may think, "He is mad, so he will hit the other guy." But in a case like this, I do think that we have to apply the general knowledge, "Angry people tend to hit other people."

So in any given case where I understand someone else, what actually is going on probably does not involve a lot of theory or a simulation. But if we want to understand what is going on, how do we learn the abilities that we are employing, I think the distinction between simulation and theory can be useful anyway, if we can distinguish *implicit forms* of them. The distinction is how I learned to do whatever it is that I do when I understand someone. If I learned it by having a similar experience myself – from the inside, from the first-person perspective – then it will be implicit use of simulation. If I learned it by observing regularities or hearing it from others or reading about it – from the outside, from the third-person perspective – then it will be a implicit use of theory.

How might this work? In some cases it is because we have had similar experiences, which we may use initially in simulations. Eventually we may not have to explicitly do the simulation but just think or a similar experience, or maybe just know how they will act because we have had similar experiences. This is *implicit simulation*. In some cases we may have learned a general rule somehow – either by observation or by instruction or by explicit reflection. These would be cases of *implicit theory*.

An advantage of thinking of both simulations and applications of theory as being usually the only implicitly employed background of an ability is that it makes it easier to see the continuity in psychological abilities from children and non-human animals to adults. Children are doing the same thing adults are doing in some of the simple cases in which they are able to understand others, but often less reliably. For example, when a 3 year-old child ascribes a true belief to someone, it seems plausible to me that they are doing the same thing I am even though I could also ascribe a false belief. It is just that I have learned new abilities on top of theirs. Similarly, children can very early have goals and employ different strategies to reach them, and remove obstacles to the realization of those goals. And it certainly seems that animals should be able do something similar, choosing between actions in order to attain a desired end, or at least monitoring and adjusting actions with respect to a desired end. And to some extent perceive others' intentions as well and coordinating their responses to them in rituals or in fighting. We can be aware of our own intentions and others' intentions without mastering all the skills that would go along with having a concept of intention or of specific intentions. Once enough of these skills are present, we are comfortable ascribing the corresponding concept to someone. The concept is not something that is invoked *in addition to* these component skills and practices but is embedded in them.

Insofar as we want to learn how these skills and practices relate to each other, and in what order they develop, thinking of simulations and uses of theory in the way I suggest could be of some use. We can think of given instances of understanding others as involving numerous abilities that all presuppose a complex learning history involving simulation and uses of theory. But in any given case where something new is learned, we can presuppose some abilities and ask what first-person experiences and what more theoretical knowledge provide the decisive resources for learning the new skill.

This could involve a child starting out with a simulation and thereby learning something new. For example, if she recognizes an adult trying to open something, she may simulate similar experiences of *making an effort* and use this to get a handle on the task at hand. As a result, she may thereby learn a new procedure for opening something, become acquainted with a new thing to open or a new object. Children learn new strategies and surprising goals this way, which enable them to imitate adults in a broader range of cases and thereby to have new experiences, which can be used in simulations, and also to new learn new goals, strategies and perceptual habits.

5.3.3 The analogy to science suggested by the concepts of theory and simulation

Since both theories suggest an analogy to science but only TT really exploits that analogy, it seems reasonable to consider what use the analogy might be in understanding ST better, and in understanding how to combine TT and ST. So we will take a look at the ways in which theory is combined with techniques, practices, experiment, traditions, etc. in constructing simulations in science and ask whether is can help us to figure out how other resources, including mental concepts and psychological generalizations, are combined with something akin to simulation in folk psychology.

In this section, I inquire whether the analogy to simulations in science can help us to appreciate the relevance of the folk psychology debate for epistemological and ontological questions about mental states. I argue that our use of simulations during cognitive development enables us to imitate the people around us and thereby to become more similar to them, which in turn makes simulation an increasingly effective epistemic strategy. Insofar as theoretical elements – such as the the beliefs and desries referred to in folk psychological discourse – play a role in imitative learning, they are *causally embedded* in our cognitive development, so we have good reason to regard them as being among the real causes of our behavior.

First, a couple of general observations. In any kind of simulation in science, theory is required in various ways, but still the simulation can tell us more than the theory alone. So, if theoretical elements have to be added to simulation theory to make it work, this does not imply a collapse of simulation theory into theory theory. And *conversely*, the analogy to science should teach us that it should be no surprise that theory theory in folk psychology needs to be enriched by something beyond theory, such as models or simulations.

Given that simulations in science involve/ presuppose some theoretical elements, a natural question to frame the discussion is: when does simulation go beyond theory? In addressing this question, I will draw upon Eric Winsberg's (2003) discussion of the ways in which simulations are more like experimental practice than theory. He makes two main points:

1) We employ a variety of techniques to derive inferences from them (e.g. data analysis, looking at results across a broad range of parameters, visualization)

2) Our judgment of the reliability of simulations depends upon our successful application of them and our successive improvement of them within practical contexts. In this connection, Winsberg refers to Ian Hacking's (1983) dictum: "Experiments have a life of their own".

The link to Hacking suggests four characteristics of the relationship between simulation and theory in science that I think can be fruitfully applied to the folk psychology discussion:

1) The construction and use of a simulation can involve numerous theories (in modeling the target system, in building the instruments). In folk psychology, much goes into making us similar enough to others to be used as models of them. For starters, evolution replaces much background theory. But this is not the whole story: In development we learn knowledge about the world, different people's roles, typical narratives, and also psychological concepts like belief and desire. All of these elements go into making our habits and dispositions similar to the people around us, but: *they do not constitute a single theory*.

2) *Theoretical elements need not be actively invoked* by the technician using a simulation. Indeed, she may not even know much about them. In folk psychology, I may use abbreviated syllogisms such as:
"Tom is angry at Paul. He is going to hit Paul."
There is no reason to insist that I must unconsciously go through a more complete derivation involving all the trappings of formal logic. I can get away with using the abbreviated form simply because the observations and inductions that justify the inference occurred during my developmental history. In fact, I can even employ true generalizations that I have learned from others without myself ever having made the requisite observations: that is a benefit of cultural evolution.

3) Theoretical insights in science may not immediately yield the local predictions we want, but are used to create models or simulations, which are then tweaked over time with the addition of new knowledge and techniques, including rules of thumb for which there is no theoretical justification (e.g. for simplifying calculations, for setting parameters) until *we learn how to use and interpret them in particular contexts*, etc. Similarly, in folk psychology

we add our increasing knowledge about the world, different people's roles, typical narratives, and also psychological concepts to our apparatus as we go along. Moreover, we have to learn to apply these elements by combining them with the right simulations. For example, it could be argued that predicting behavior on the basis of a false belief involves a concept of belief plus a simulation. The idea would be that once I have figured out what beliefs someone has, the only (or the usual) way to get from the beliefs (in combination with desires and much else of course) to a prediction of action is to imagine I had those beliefs and see what I would do.

4) *Applying simulations helps us to gain theoretical insights which, in turn, help us to improve the simulation system*. In folk psychology, one can draw a useful analogy to children's imitative learning. From a very early age, children imitate novel movements and even actions – often with less than complete understanding of the intention guiding the action (Metzoff 2005, Tomasello 1999). In so doing, they learn new actions and become acquainted with new intentions (i.e. goals and strategies for achieving those goals). As I mentioned in discussing Tomasello in section 5.1.1.2, children begin around nine months to show signs of a more sophisticated understanding of their own agency, and iot is then that their imitative behavior increases dramatically (Tomasello 1999). This suggests that they are using their own recalled experiences of performing actions as a heuristic basis with which they gain a foothold upon adults' actions. But the interpretation of this data is still controversial. At any rate, through imitation they learn the habits and dispositions that are typical of the members of their culture, which leads to their being cognitively structured in a way similar to the people around them. The result of this is twofold: they can explain and predict others' behavior more effectively by simulating them, and their behavior can in turn be explained and predicted simulatively by other members of their culture.

5.4 A look at some hybrid approaches

Shaun Nichols, Ian Ravenscroft, Steven Stich and Josef Perner (Nichols and Stich 1995, 2003, Stich and Ravenscroft 1994, Ravenscroft 2008, Perner 1996) have done a great deal of work toward articulating the theoretical option of TT-centrist hybridization. I will first dicuss their reflections (5.3.3.1), then move on to Goldman (2006)'s ST-centrist hybridization (5.3.3.2), and finally to some neuroscientific approaches that are also ST-centrist (5.3.3.3).

5.4.1 TT-centrist hybrids

As Ravenscroft (2008) points out, it is possible to distinguish between what he calls external and internal consturals of both TT and ST. Let's look at how this could work. Starting with the external construal, either TT or ST could be regarded as an abstract or functional account of the processes underlying folk psychology.

TT, construed externally, would be saying that the most parsimonius way to *describe* folk psychology in terms of ascription of mental states and predicitons on the basis of psychological generalizations, but what actually is going on in the brain is that (as ST would have it) we are undergoing mental states and processes in an "offline" or "pretend" mode, and using these simulations to predcit and/or understand others. Simulation could be said, in this event, to "implement" theory. If one is internal with respect to the other, we could say that it is primary in sense #5 formulated in 5.3.1.

Alternatively, one may construe TT in an internalist sense. Folk psychology would then be an "internaly represented knowledge structure used by the cognitive mechanism underlying our folk psychological capacities. In this event, one could describe the achievement of such a cognitive system by saying that we put ourselves in others' shoes and simulate them, but this is not doing any explanatory work.

Stich and Nichols (1992) also consider way in which ST could be implemented by TT but still do some explanatory work. Specifically, it could be the case that we indeed run simulations in predicting others' behavior in the sense that we use our own decision-making resources or practical reasoning), but that these resources turn out to depend upon a implicit decision theory (47). Simulation, in this event, could be implemented by theory with respect to decision-making, and yet not be implemented by theory with respect to the selection of pretend mental states that form the input to the decision-making procedure.

Perner (1996) goes still further in integrating ST into TT. For the same reasons as Heal, he does not think it likely that we use a theory of rationality to ascribe rationality to others. Hence, Perner thinks, with respect to predicting what other people will deem rational, we use our own procedures for deciding what is rational or not. According to Perner, the same goes for lots of different kinds of judgements that we make. Think of grammaticality, for example: if asked to predict whether some other speaker of my language will judge a sense to be grammatically correct, it seems likely that I just check whether I would judge the sentence to be grammatically correct and expect the same of the other person rather than applying some sort of theory of sentence comprehension in deriving my prediction about their judgments.

Perner calls this sort of first-person knowledge “predicate-implicit”, since we only implicitly predicate it to ourelves. In other words, we do not represent it as being our own conception of rationality or grammaticality or whatever, decide that the target person is similar to us, and and then attribute it to them. Rather, it just appears to us unproblematically that a certain judgment is objectively correct. Implicitly, we are appealing to our own conception of rationality, grammaticality, etc. In short, Perner goes beyond Stich, Nichols, Ravenscroft in allowing for ST to actually do some explanatory work. In these cases, simulation would be primary in sense #3, i.e. in the sense of a default strategy, since we would presumably turn to other explanations if the target person’s judgment turned out not to match ours.

5.4.2 Goldman 2006 (An ST-centrist hybrid)

In his most recent book, Goldman (2006) has advocated the view that the best theory of folk psyhcology will probably turn out to be a hybrid. One important kind of case in which simulation and theory could complement each other is in what Goldman calls retrodiction (in contrast to prediction). In retrodiction, you want to ascribe someone an intention or a set of beliefs and desires on the basis of observed actions. Goldman ascknowledges that a straightforward simulation is not directly applicable in such a case, since doing a simulation (in his view) requires one to start out by selecting pretend mental states, on the basis of which one can simulate the choice of an action. Goldman’s claim (or consession, depending on your perspective) is that simulation could not be of any assistance in selecting those input states, sine the input states precede and indeed cause the action which one is observing. Simulatively interpreting the input states on the basis of the observed actions would therefore require simulating in reverse! And since, as Goldman says, psychological processes do not likely run in reverse, it is unlikely that we use simulation to interpret the input states (i.e. the other person’s beleifs, desires, intentions, etc.) on the basis of observed behavior (2006, 45).

His suggestion is that we might use theory to construct hypotheses about what intentions, beleifs and desires the target person might have. Specifically, we could use generalizations about what beliefs and desires people have in certain situations in order to select the pretend mental states. Then we would run a simulation to see how we would act given the hypothetically ascribed beliefs and desires. If the action we would produce matches the action we observe the other person performing, the hypothesis is confirmed and the ascription is accepted. If not, we repeat the process until we get a match (2006, 45). On this model, simulation would be the endpoint of the explanatory project and would therefore be primary sense #4 mentioned in section 5.3.1.

There is an option for simualtion theorists who want to resist this line of thought. They could counter that one could simulatively construct hypotheses by observing the situation rather thatn the target person's behavior, and thus simulate the target person's *perception*. One could thereby arrive at hypotheses about their beliefs and desires and test them out in the manner sketched by Goldman. Nevertheless, Goldman's proposal certainly offers us at least and interesting model in which theory and simulation could plausibly complement each other.

5.4.3 Neuroimaging-based ST-centrist hybrids

Another empirically motivated alternative view that departs from the conception of theory of mind as a single unified ability is a suggestion made by Helen Tager-Flusberg and Kate Sullivan (2000). They break up theory of mind skills into two separable components, a social-perceptual and a social-cognitive component, and present neurobiological, developmental and pathological/clinical evidence supporting this division. What they seem to have in mind is basically a hierarchical conception according to which simulative processes constitute the basis of theory of mind skills in general, while meta-representational skills along the lines of theory theory are built on top of these more basic skills.

The social perceptual component, which they link to the amygdala and some regions of the medial temporal cortex, is especially involved in identifying emotional expressions. In addition, they note that it is involved in attributing "other person-related knowledge (such as personality traits) primarily on the basis of immediately available perceptual information" (Tager-Flusberg and Sullivan, 2000, 62). This is a bit vague, but they do give a few hints about what social stimuli besides emotions can be perceived with this component. For example, they go on to mention "perception of biological or intentional motion" (63)[62], as well as identifying people's race or gender by looking at them. What they also appear to have in mind is the idea that there is a range of relatively simple cases where we can understand others' intentions simply by mirroring them – that is, by entering into a state that approximates the state they are in. In their view, this can occur without employing metarepresentational skills, explicitly ascribing a state to them, or, as they put it, "reasoning

[62] The amygdala has been linked to perception of biological and intentional movement by, for example, Bonda, Petrides, Ostry and Evans, 1996)

about behavior". This is not a trivial point. One could be inclined to think that perceptual processes play a role in action understanding, and even that they constitute a distinct component, but stop short of saying that they ever, even in simple cases, suffice for understanding socially significant stimuli. In other words, one could regard them as a necessary component in some cases, but never as sufficient. But Tager-Flusberg and Sullivan apparently regard them as necessary and sufficient in a certain range of cases.

As for the social-cognitive component, Tager-Flusberg and Sullivan identify it with the "conceptual understanding of the mind as a representational system" (61) and note that it is linked with other cognitive capacities such as "theory-building", language in general and syntactic skills such as sentential complements in particular - i.e., the kinds of things that theory theorists like to talk about. They point to the developmental synchrony of such skills, which have in common that they may plausibly involve metarepresentation (Carlson, Moses and Hix, 1998; Hughes, 1998; Roth and Leslie, 1998). The neural substrate they propose involves regions of the prefrontal cortex, for example the orbital frontal cortex.

The distinction is supported by various studies comparing different specific pathologies, i.e. picking out differential disabilities in children with Williams syndrome (WMS), Asperger syndrome[63] and autism spectral disorder (ATS). Specifically, they propose that the social-cognitive component is impaired in people with WMS, the social perceptual part is impaired in Asperger syndrome, and both are impaired in autism. People with WMS apparently perform pretty well at identifying emotional facial expressions, but not so well at tasks in which specifically meta-representational skills may plausibly be said to be required, such as false belief tasks[64].

People with Asperger Syndrome, in contrast, are relatively poor at attributing social significance to ambiguous visual stimuli (Klin, Schultz and Cohen 2000), interpreting mental states as expressed in eyes (Baron-Cohen et al., 1997) and matching emotion words to emotional facial expressions (Baron-Cohen et al., 1999). It is telling that they do not activate the amygdala during this kind of task, whereas normal individuals do. On the other hand, they perform significantly better than autists at higher-order theory of mind tasks. Interestingly, though, they do not activate the same regions of the medial frontal cortex as normal individuals when performing these tasks (Happé et al., 1996). This suggests that they are employing alternative resources in the absence of the typical mechanisms. The question

[63] Asperger syndrome is generally regarded as a mild form of autism.

[64] Of course, they do not perform as well as normal individuals. To say that they perform well in this case means that they perform better than other individuals who have mild to moderate mental retardation for other reasons, and therefore have similar IQs but different cognitive profiles. Here, they were compared with groups of individuals with Prader-Willi syndrome.

obviously arises what alternative resources might be involved here. Tager-Flusberg and Sullivan point out that their success at these tasks is correlated with their syntactic skills (Tager-Flusberger and Sullivan 2000), and suggest that they are making the most of linguistic resources to compensate for the absence or weakness of other representational or metarepresentational skills that would normally receive input from the perceptual component, e.g. from the amygdala, but cannot do so because this component is impaired. In evaluating this proposal, it would be interesting to know whether areas especially involved in language production or comprehension – such as Broca's area – are abnormally active in these individuals during these tasks. It would also be interesting to know how well the linguistic skills that they tested correlate with proficiency in the emotion-recognition tasks in normal individuals.

Tager-Flusberg and Sullivan's proposal has since been taken up and supported by empirical studies conducted by Pineda and Hecht (2008), and similar ideas have been put forth by Keysers and Gazzola (2007). The upshot of these studies is that there appears to be a distinction between two different brain circuits for social cognition, one comprising posterior and middle insula, ventromedial prefrontal and premotor cortex, and one comprising midline structures and tempoparietal junction and anterior insula. The proposal is that the former circuit, which is more prominently used in emotion recognition and in understanding simple intentions that are transparent on the basis of observed bodily movements and perceptible features of situations, is best characterized as functioning by simulation, whereas the latter, which is prominent when these movements and features have to be interpreted, reported verbally, or understood in light of information that is not perceptually accessible, is best characterized as employing more reflective, linguistic, hence theoretical, processes.

Indeed, it is a very plausible proposal in light of our analyses of TT and ST. Simulation would work insofar as we are similar to others, and should be (evolutionarily) preferred in such cases because it would be parsimonious. Insofar as our difference from others increases, theoretical elements should become increasingly useful, thus overcoming the pull toward parsimony. And we would expect that our difference to others should increase as situations become more complex, requiring for their interpretation more information, and more multifarious information that is not bounded to the perceptible situation. These models present simulation as being more or less a default strategy, and therefore primary in sense #3 (see section 5.3.1).

5.5 Taking stock

We have seen that there is at present no clear way to decide empirically between TT and ST, and that both sides (or, rather, all of the various versions on both sides) have certain theoretical virtues in their favor. We have also seen that this situation has led some theorists to favor dropping both theories and starting fresh. Although this impulse is understandable, I think it is worth trying to make the most of all theoretical reources that are currently available, and TT and ST offer such resources. I therefore favor the approach of looking for ways to combine theoretical and simulationist elements. This approach certainly does not exclude the option of making use of ideas put forth by Gallagher, Zahavai and others who take different approaches.

It seems reasonable to me, as a rough working hypothesis, to suppose that simulation processes would be used whenever they are likely to work, and that theoretical elements would come into the picture when simulation is less likely to work. This preference for simulation as primary (in particular in sense #3 in section 5.3.1, i.e. in the sense of a default strategy) is based upon empirical and theoretical points: empirically, ST yields novel predictions that have been supported by empirical work, in particular on mirror neurons (chap. 6) but also on concepts and action representaiton and planning (chap. 7 & 8). Theoretically, these same empirical findings suggest very plausble ways in which we can understand theorizing or simulating as implicit processes, and that is exactly what both TT and ST are in need of, especially in light of the various common-sense or phenomenological challenges to the the idea that we explicitly theorize or simulate. Moreover, simulation seems to be evolutionarily more parsimonious. Since simulation seems likely to work should work best when we are similar to the target person. It is likely to be supplemented, corrected, or replaced by theory when a target person is likely to be relevantly different.

My strategy is to start out with what I take to be the basic insight of ST – that in making sense of others' behavior *we undergo the same procedures that we would undergo if we ourselves were deciding upon, planning or executing an action in the same circumstances* – and ask what can be added to this picture to account for social cognition in various contexts (third-person, interactive) and with different kinds of motivation (individual, helping, sharing). In the rest of this dissertation, I will be attempting develop this approach further, and will in so doing aim to make the concept of simulation more specific – ideally in in a way that reflects and also heuristically benefits ongoing research. Hence, we will be looking at the way in which the concept of simulation is used in various contexts in neuroscience, cognitive science and psychology, and attempting to relate these usages to ST in its various versions that we have discussed so far.

In a sense, then, the second half of this dissertation breaks with the first part in that we will no longer be comparing two theories of folk psychology but trying to improve upon the insight introduced by one of those theories. In doing so, I will go into some detail concerning the empirical work that supports ST, and which ST should be modified in order to accommodate. That will lead us, first of all (chap. 6), to look at research on mirror neurons (MNs), which has provided powerful experimental support for ST. We will see that the concept of simulation needs to be refined in order to exploit this experimental support and to link ST with MN research. Since simulationist theories of concepts suggest an appropriate way of refining the concept of simulation, chapter 7 will be devoted to theories of concepts. Chapter 8 will be devoted to particular issues that need to be addressed in applying this refined concept of simulation to mental concepts and to ascription, which are necessary for action understanding and thus also for a theoretical account of folk psychology.

Part II: Improving on simulation theory in light of recent empirical findings

Chapter 6:
Mirror Neurons and Simulation Theory

6.0 Introduction

In this chapter I will review recent empirical work and theoretical interpretations of mirror neuron (MN) research, which has been claimed to offer support for ST. I conclude that this claim is correct, but that it is unclear which version of ST is supported. Resolving this issue will require further empirical work, but also theoretical refinement of the concept of simulation as well as of ST itself. Hence, making use of this empirical support is a two-way street: it will require making ST more precise in a way that fits with and contributes to the ongoing research.

6.1 what are mirror neurons?

So-called 'mirror neurons' have been a sensation since their discovery in the early 1990's.[65] The term 'mirror neurons' was introduced by Giaccomo Rizzolatti to designate neurons in the F5 area of the brain of the macaque monkeys that have the surprising property of firing when the animal observes as an action is being performed. That is surprising because the F5 area is one of two distinct areas (the other being F4) making up the ventral premotor cortex in macaques[66], which is involved in the preparation of *action*, i.e. it would not seem to have anything to do with *perception*. In fact, the neurons in F5 that fire during action observation are not just out-of-place perceptual neurons; they fire while the animal is either performing an action or observing the same or a similar action. The catchy term 'mirror neurons' is meant to highlight this symmetrical property. It is important to note that only a subset (approximately one third[67]) of MNs are *strictly congruent*, i.e. fire when one particular action is performed or

[65] Di Pellegrino et al., 1992, Rizzolatti et al., 1996
[66] Matelli, M., Luppino, G. and Rizzolatti, G., 1985.
[67] Rizzolatti and Sinagaglia 2008

observed and otherwise never. We will come back to this point during the course of this chapter.

Since measuring the activity of individual F5 neurons involves sticking electrodes into the macaques' brains, there are ethical hindrances to performing the same kind of study in human brains. Hence, most of the studies of MNs I will be mentioning in this chapter were conducted with apes, and a skeptical reader may well think that all this data from studies with apes does not tell us much of anything about humans. Although this is a valid point and should make us hesitant about boldly drawing spectacular conclusions, it would not be justified to dismiss all of this data as irrelevant to understanding humans. Although we cannot usually record the activity of individual neurons in humans, there are also lots of studies with humans in which the measurement of neural activation in fairly specific areas very strongly suggests that humans have a similar mirror system. In order to convince the skeptical reader to keep an open mind, I will start out with a brief overview (section 6.2) of some of the different methods that are used with humans, pointing out their potential benefits and limitations. Since I will be dealing with particular studies more closely during the course of the chapter, I will not go into much detail at the outset. I only want to convince the skeptical reader that there is enough data from studies with humans to show that there are at least interesting similarities between humans and apes with respect to MNs. Hence, studies on apes are relevant to understanding the human MN system.

Since the primary aim in this chapter is to assess the relevance of MNs for action understanding and in particular for the simulation theory of action understanding, I will then (section 6.3) present the conception proposed by Goldman and Gallese (1998), which constituted the first simulationist interpretation of MNs and initiated a discussion that is still ongoing and has led to the further precision of ST and of theoretical conceptions of MNs. In doing so, I will discuss a few major points of criticism that have been raised, as well as responses that have been formulated. Then I will discuss subsequent theoretical developments (6.4). In section 6.5, I will discuss some versions grant MNs a subordinate role in action understanding that is dependant upon the ascription of an agent's prior intention elsewhere in the brain. Goldman, for example, presently espouses a position of this kind. Other versions, which will be the topic of section 6.6, are more radical in that they seek to do without any distinct metarepresentation that would constitute an ascription. Gallese is now the most prominent proponent of this approach, and I will devote special attention to clarifying his position (which suffers a bit from lack of clarity). In section 6.7, I express some lingering concerns about robust interpretations of MNs in action understanding. In section 6.8, I consider the relevance of MNs for other versions of ST (Gordon, Heal). The conclusion of the chapter will be that MNs provide empirical support for ST as against TT but leave the choice between versions of ST is open. Moreover, this empirical support does not come for free; it is an invitation and a challenge to refine ST in order to make the most of it. This challenge will be taken up subsequent chapters, in which I will be considering simulationist theories of concepts (Barsalou, Lakoff, Prinz) and of action representation and planning (Jeannerod, Pacherie), which modify the notion of simulation in a way that helps to tie together Goldman's and Gordon's versions of ST.

6.3 MNs in humans:

6.3.1 Methods

The first kind of study carried out with humans makes use of the electroencephalography (EEG). The EEG records cortical electric activities, the rhythms of which are classified by wave frequency. One of these rhythms, the so-called mu rhythm, is of special interest in connection with MNs. The designation is used to refer to amplitude oscillations in the 8-13 Hz range, which prevails in central regions of the brain when the motor system is at rest, but is desychronized during motor activity. As early as 1954, it was found that the mu rhythm is desychronized not only during the performance of motor activity, but also during observations of others' motor activity (Gestaut and Bert 1954). These studies have of course been repeated more recently, with similar results. During performance of actions, of course, MNs are not the only neurons active in their vicinity, and so the suppression of mu rhythms during performance of actions cannot be uniquely attributed to MNs. Rather, other active neuronal systems in the motor, premotor, sensori-motor cortices could equally be involved. But during action observation, MNs are the only neurons in the area that are active. Hence, it seems plausible to suppose that their suppression could be used as a measure of MN activity when other neighboring areas are not active, i.e. during action observation. There are also some other findings that support this suggestion. (Oberman et al 2007: 191). This presents us with a relatively cheap and non-invasive method for investigating the involvement of MNs in particular abilities – such as social cognition, which is our interest here. I will come back to them in section 6.2.2.

Magnetoencephelography (MEG) studies, which analyze the electric activity of the brain with the help of recordings of the magnetic fields produced by this activity, have also provided evidence that mu rhythms in the precentral cortex are desynchronized during performance or observation of motor activity.[68] A major drawback of EEG and MEG studies is that they do not precisely localize the cortical areas and neural circuits involved in the activity they record.

Transcranial stimulation studies (TMS) have also yielded corroborating evidence. In TMS studies, a magnetic field produced by a coil held close to the head induces an electrical current in the motor cortex. This current makes it possible to record the motor evoked potentials (MEPs) in the contralateral muscles. Fadiga et al (1995) used this technique to record MEPs in the muscles in the hands of subjects observing agents performing actions. That the MEPs increased during observation of transitive actions was coherent with studies on monkeys. It was surprising, however, that they also increased during observation of intransitive actions. TMS studies are also limited in that they do not enable very precise localization of cortical areas and neural circuits (Rizzolatti 2008, 118).

Positron emission tomography (PET) and functional magnetic resonance imaging (FMRI) studies, both of which record blood flow in the brain, complement these other techniques well in that they enable greater localization of activity in the brain. Studies using fMRI have confirmed that the activity of the human mirror system is located primarily in Brodmann's area 44 (Rizzolatti 2008, Buccino 2001), which is the posterior part of Broca's area and is the human homologue of F5 in the macaque (Petrides and Pandya 1997). In addition, the human mirror system appears to include portions of the premotor cortex and the inferior parietal lobule (Rizzolatti 2008).

FMRI studies have also confirmed the finding, mentioned above, that the human mirror system is active during the performance of intransitive actions (Buccino et al 2001). In this study, human subjects watched videos in which a human, a monkey or a dog performed transitive or intransitive actions (the transitive action was grasping a piece of food with the mouth; the intransitive action was silent speech). Interestingly, MN activity was greater during observation of human speech than during observation of a monkey smacking its lips (a communicative gesture) and absent during observation of a barking dog. Buccino concludes that MNs are not active when the action being observed is not in humans' motor repertoire.

Despite their advantages in localization, fMRI studies also have certain drawbacks – blood flow only indirectly reflects the activity that we are actually interested in, and its significance is difficult to interpret (Logothetis 2008, 869-70). Moreover, even the activity that it indirectly reflects is not unambiguous – it could, for example, reflect either the

[68] Hari 1998

functioning of a particular neural circuit or the inhibition of this functioning Logothetis (2008, 872).

In short, although experimenters are more limited in their investigation of human subjects, the combination of these various techniques – not to mention the phylogenetic proximity of apes to humans – makes the case very strong that humans have a mirror system similar to that of apes, but with greater sensitivity to intransitive actions.

6.2.2 MNs in humans: evidence derived by measuring suppression of mu rhythms

I mentioned in section 6.2.1 that mu rhythms are suppressed during performance and also observation of actions, and that their suppression is regarded as a measure of the activity of MNs. There has been some highly interesting empirical work using this basic idea, which links MNs to social cognition. The first group of studies deals with pathological cases, namely with autism, while the second group of studies involved using deliberate manipulation of mu rhythms to influence social cognitive abilities.

Since autistic individuals show impaired social cognitive abilities like theory of mind, joint attention, empathy and language, one might suppose that their mu rhythms are not suppressed during action observation. Oberman et al (2005) therefore conducted an EEG study in which they found that mu rhythms in autistic individuals are suppressed during performance of particular actions using their hands but not during observation of the same actions, whereas mu rhythms in non-autistic individuals were are suppressed both during performance and during observation of the actions. Since the suppression of mu rhythms is associate with the activity of MNs in action observation, this result suggests that the MN system in autistic individuals is impaired, and thus also that the MN system is involved in the social cognitive abilities at which austistic individuals fair poorly.

We should of course be cautious. The finding does not demonstrate just how MNs are involved in social cognition. In fact, it is perfectly compatible with the thesis that MNs play a subordinate role in social cognition that is dependent upon the functioning of a theory of mind in the sense of theory theory. We just have to imagine that there is theory of mind module or something of the kind that is responsible for interpreting the observed movement as an action, but which in autists is not fully functional, and therefore does not interpret observed movements readily as actions, and that MNs are therefore not activated.

But there is also another strand of research involving mu rhythms that corroborates the claim of a more active role for MNs in action understanding. Specifically, the connection between autism and MNs (via mu rhythms) has raised the question whether social cognition in autistic individuals could perhaps be improved by influencing their MN activity, possibly via mu rhythms. Pineda et al (2008) have therefore conducted studies involving neurofeedback training (NFT), which are designed to enable participants "to learn to self-regulate endogenous brain rhythms" and thereby "to access and control regulatory systems that increase/decrease synchronous activity in neuronal cell populations." (Pineda 2008, p. 3) In one such study, which consisted in 15 hours of training over the course of ten weeks, autistic individuals in the experimental group played a variety of video games (involving race cars and the like). On the left-hand side of the screen there was a bar displaying the level of activation in the 8-13 Hz range (mu rhythms), while on the right-hand side there was a bar displaying activation in the 30-60 Hz range. The participants were able to advance in the game only when the bar displaying mu rhythms was above a threshold (which was progressively raised during the training) and the bar displaying the other, non-mu, activity was below a certain threshold.

When the training was completed, the participants were assessed in a battery of tests of social cognitive abilities and their scores compared with their performance on the same tests before the training. They were also compared with corresponding tests upon members of a placebo group, which was trained in a similar manner, except that the bar supposedly displaying their mu rhythm activation did not really reflect their mu rhythm activation. The result was that only the members of the experimental group had improved at a visual form of the test of variables of attention (TOVA), which measures sustained attention ability, and in various subscales[69] of the Autism Treatment Evaluation Checklist (ATEC). Both groups improved at a test measuring their ability to imitate hand, arm and finger movements[70]. Mu suppression during action observation also improved markedly in the experimental group (and not in the placebo group).

6.3 What use are mirror neurons to simulation theory?

Although there are numerous versions of simulation theory, it is Alvin Goldman who has most actively brought the discussion of mirror neurons to bear upon simulation theory. I will therefore start out from the position espoused in Goldman and Gallese (1998). Later in the chapter, we will see that Gallese and Goldman differ in their further development of that position in response to criticism.

[69] The ATEC is a questionnaire that includes four subscores: speech/language communication, sociability, sensory-cognitive awareness, and health/physical behavior. Pineda 2008, p.6

[70] specifically, the apraxia imitation scale

6.3.1 Goldman and Gallese

Gallese and Goldman (1998) were the first to claim that the discovery of mirror neurons (MNs) provided empirical support for simulation theory (ST) as opposed to theory theory (TT) as an account of folk psychology. According to the version of ST they defend in that paper, ST agrees with TT in postulating that folk psychologizers represent the mental states of the people whose behavior they are interpreting (i.e. prediction or retrodiction), but differs from TT in denying that folk psychologizers represent, even implicitly, the psychological laws that link mental states to each other and to input and output.

I just want to note at this point that Gordon's version of ST would differ from TT on both of these points, since, for him, simulation is supposed to save the psychologizer the trouble of representing the other person's mental state – or at least the trouble of representing *it as their mental state*, i.e. of ascribing it to them, which would qualify as a meta-representation. It will become clear later on that what Gordon has in mind in fact does qualify as a representation in Goldman's terms. Nevertheless, this match is misleading, since they have different conceptions of mental representation and thus also meta-representation and ascription. For Gordon, the intention being ascribed would not be introspectively accessible. Presumably, if and when the folk psychologizer became aware of it (i.e. in giving a verbal explanation), it would be an extrapolation from the prediction of behavior that she came to without using mental concepts[71].

For Goldman, on the other hand, the ascription of an intention cannot be merely extrapolated from a prediction, since the prediction must proceed from the ascription. Given that his version of ST is meant to dispense with the need for psychological laws in folk psychology, the mental states that are ascribed to others and to oneself cannot be defined or differentiated via their nomological connections to each other and to behavior, as all versions of TT assume. Indeed, as we have already seen in our discussion of Goldman's version of ST, Goldman draws the consequence that there must be some kind of introspective access to mental states, and therefore joins his ST with a critique of functionalism.

Interestingly, Gallese and Goldman have drifted apart since 1998, and now Gallese seems closer to Gordon in his understanding of representation of other's intentions. In any case, what they all have in common is the conviction that folk psychologizers do not need psychological laws to get from mental states to actions, since they undergo the same mental

[71] She would make the prediction in the context of a simulation and then derive a verbal report by using the ascent routine, which does not presuppose full mastery of the relevant mental concept (see 4.3).

processes as the person whose behavior is to be interpreted and thereby wind up with an expectation of the other's behavior that is at least as good as they would get by using psychological laws.

For the time being, I will turn back to Goldman and Gallese (1998). The simulation heuristic, on their view, can be used either predictively or retrodictively. For *prediction*, one has to start out by ascribing a goal and a constellation of beliefs to the target person, and then pretend to have this goal and these beliefs, and choose an appropriate action to pursue. Finally, one takes the output of the simulation as a prediction of the target person's behavior. In *retrodiction*, as I discussed in section 5.3.3.2, one starts out by observing the target person's behavior. One then conjectures what goal the person might be pursuing, pretends to have that same goal, and then simulates the procedure whereby an appropriate action is chosen. If the action one chooses matches the behavior of the other person that one is observing, one concludes that the other person did indeed have the conjectured goal. In both cases, using one's own decision-making procedure to get from a goal to an action therefore obviates the need to represent the laws or principles employed by the target person.

According to Goldman and Gallese (1998), studies of MNs provide evidence of processes or events in observers' brains that match processes or events in agents' brains, and could therefore play the role of simulation heuristics. Since TT predicts no such matching, they assert that MNs provide empirical support for ST. They are careful to note that MNs are not likely the whole story about how the simulation heuristic is instantiated. More likely, they think, MNs constitute a primitive version upon which humans' more sophisticated simulation heuristic is built.

6.4 Criticism of Goldman and Gallese (1998) and similar conceptions

6.4.1 Criticism 1

But some critics (Jacob, Pacherie, Csibra) object that Goldman and Gallese fail to distinguish adequately between bodily movement or motor goals, on the one hand, and action in pursuance of prior intentions on the other. It is hard to see how the motor system alone could suffice to ascribe an intention, as would seem to be essential to action understanding, since one and the same movement can constitute the means of carrying out diverse actions, and one the same action can be pursued by means of various movements.

Response to Criticism 1

Experimenters have come up with some interesting responses to this objection. For one thing, most F5 neurons respond when the monkey observes an action upon an object, but not if the same bodily movement is observed in the absence of the target object (Rizzolatti and Sinigaglia 2008). This suggests that the MN activity does not represent mere movement but something more complex, arguably action. Umiltà and colleagues (2001) conducted a study which developed this point a bit further. In this study, the monkeys observe as an object is placed behind an occluder, and then observe the beginning of a grasping action conducted upon the occluded object, but do not see its completion – i.e. the contact with the object occurs behind the occluder. MN activity during observation of the incomplete actions turned out to match that during observation of complete actions.

Another compelling group of studies has found individual bimodal (audio-visual) MNs that are active during performance of a particular sound-producing action (such as breaking peanuts), and also when the monkey sees or hears the action being performed (Kohler et al 2002). This is a fairly clear demonstration that the MNs are not sensitive to mere movement but to something more abstract, arguably actions. But it still falls short of demonstrating that the MNs achieve action understanding, i.e. that they either cause or constitute ascription of a prio intention. Ascription of a prior intention could occur elsewhere on the basis of diverse perceptual information, and the MN activity could reflect this, rather than causing or constituting it.

Iacobini et al (2005) also report a clever study intended to answer to the same objection. To whit, they believe they can "show that the mirror neuron system also plays a role in coding the global intention of the actor performing a given motor act." (Iacobini et al 2005, 0529) In their study, there are three conditions. In the context condition, subjects see a scene with cups, mugs and plates arranged as if either before tea (intention to drink) or after tea (intention to clean). In the action condition, they see a human hand grasp a mug. In the intention condition, they see the hand grasp the mug within one of the two context conditions. The result was that viewing the hand grasp within the intention to drink context produced a greater increase in MN activity vis-à-vis the action condition alone (just the grasping hand) and the context condition alone (objects arranged as if before tea), than viewing the same hand grasp within the intention to clean context vis-à-vis the action context alone (the grasping hand) and the context condition alone (objects arranged as if after tea).

They argue that the differential activation must be the result of the ascription of two different intentions, since the perceptual data remains constant, and conclude that "this mirror neuron area actively participates in understanding the intentions behind the observed actions." (Iacobini et al 2005, 0532). What they mean by "actively participates" is somewhat ambiguous. Sometimes they say the MN system "plays a role" in action understanding (0529), and sometimes they seem to interpret it more robustly, i.e.: "to ascribe an intention is to infer a forthcoming new goal, and this is an operation that the motor system does automatically." (0529)

It is unclear whether Iacobini et al's response to Jacob's challenge works. There is reason to doubt even the weaker interpretation of Iacobini et al's thesis. Jacob points out that even if the differential activity of MNs in this study is taken to reflect that different intentions are attributed in the different scenarios, it does not prove that MNs are the origin of these attributions. Rather, perceptual cues could lead the subjects to ascribe different intentions, which could then be the cause of differential MN activity (Jacob 2008, 210).

6.4.2 Criticism 2

Csibra (2005) offers a powerful conceptual argument to the effect that action understanding cannot derive from motor simulation. He starts out from empirical results reported by Gallese et al (1996). In this study, some neurons in the F5 area of the premotor cortex in macaque monkeys are active when performing or when observing an action whereby food is grasped. Interestingly, no MNs are activated when the grasping action is observed in the absence of the target object (the food). Why not? Apparently because the action has no goal and therefore no meaning to be understood. But this implies that the action is interpreted (a goal or outcome identified) before the MNs enter the game, which is inconsistent with Gallese's view that mirror neurons play a key role on interpreting actions, i.e. identifying the goal or outcome.

Response to criticism 2

The response that seems most promising is to concede that the recognition that an action is taking place occurs in some other area(s) of the brain, and that the MNs serve the function of identifying more specifically what the action is. Susan Hurley takes this line:

> The conclusion does not follow that action understanding cannot be grounded in motor simulation. The information that an action is directed to a particular goal is richer than the information simply that an action is goal directed. Even if motor simulation does not provide the latter information, it can provide the former; this provides a good sense in which understanding the specific goals of observed actions may be enabled by motor simulation. (Hurley 2005).

Csibra finds this idea unconvincing. He does not think it makes sense for some other area of the brain to recognize the movement as goal-directed and therefore as an action without recognizing toward what goal it is directed. "This would be akin to recognizing that an object is 'bigger than something' without specifying the 'something'" (Csibra 2005). This counterargument seems pretty persuasive to me. It seems likely that visual areas (perhaps in the STS) must identify the movement and the object and determine that the combination of the two constitutes an action and that MNs only then get involved.

We will see below that Csibra thinks they contribute to predicting ongoing behavior on the basis of an ascription of a prior intention, but not that they actually contribute to understanding actions currently being observed. The question at the moment is whether the MN activity can be taken to contribute to action understanding. One possibility: it would be conceivable that the visual areas could identify the movement and the object, and determine that the combination likely constitutes an action. The information that such-and-such a movement and such-and-such an object are being observed in such-and-such positions would then be sent off to the MNs. If there are any MNs that are sensitive to that constellation of representations, they too would become active, thereby confirming that an action is taking place. The MNs would then become active and simulate various actions until one matched the observed action.

Another possibility, close to Hurley's proposal, is that the action is identified in a thin manner in the visual areas, and that the activity of the MNs enriches the representation. But, as I mentioned, I do not think they can enrich it by adding the specific goal, as Hurley suggests. Rather, they could enrich it by introducing an experiential aspect. This would in itself constitute a different kind of understanding, and would also enable one to access further interpretive resources, e.g. by triggering associations with similar actions one has performed oneself. Finally, in line with Csibra's assertion that MNs are predictive, it would improve one's ability to predict the ongoing action of the agent being observed.

The upshot is that Csibra's objection makes it seem highly likely that other areas of the brain are involved in identifying and understanding actions. MN activity may either follow upon the prior achievement of action understanding, or it may participate in or enrich it.

6.4.3 Criticism 3

This criticism is based on the fact, mentioned above (intro to this chapter) that only a subset of MNs is strictly congruent. Strictly congruent MNs fire when observing or performing one and the same action (same type of grasp and same object).

Many other MNs are responsive to multiple actions. They may be active during the execution of only one action but active during the observation of several actions, or active during the execution of several actions but to the observation of only one action. Obviously, there are some fuzziness about how to differentiate actions, so it is no surprise that the estimates of the proportion of all MNs that are sensitive to multiple actions differs greatly among researchers: estimates range from 21 % (di Pellegrino et al., 1992), to 33 % (ingestive actions in Ferrari et al., 2003), 37 % (PF MNs, Gallese et al., 2002), about 40 % (Rizzolatti et al., 1996), 45 % (Gallese et al., 1996), or even 68 % (Umiltà et al., 2001). Csibra notes, then, that a neuron that is associated with the 'grasping by hand' motor action could be activated by the observation of 'hands interaction', or by 'grasping with the mouth' (Gallese et al., 1996). Taking the idea of simulation seriously, according to Csibra, “these instances of MN activation should be classified as mis-simulation or mis-interpretation of the observed action. As the high proportion of MNs responding to multiple actions suggests, mis-interpretation of observed actions would not be exceptional.” (Csibra 2005).

Beyond this, many MNs are fire when one action is executed or when a *functionally related* action is observed. "For example, the effective observed action was placing an object on the table, whereas the effective executed action was bringing food to the mouth or grasping the object" (di Pellegrino et al., 1992, p. 179). Csibra thinks this counts against the utility of MNs in simulation. He notes, “if mirror neurons implemented a simulation procedure, this example would literally mean that the monkey understood the object-placing action as having the same meaning as when he grasps (or eats) an object.” So these cases would also constitute, in effect, failed simulations. Altogether, MNs that are active during the performance of an action and during observation of a related action or of multiple actions makes up something like 60% (Fogassi & Gallese, 2002) or 70% Rizzolatti and Sinagaglia 2008, 84). of all MNs. Taken together, they constitute the class of “broadly congruent” MNs.

The upshot of Csibra’s criticism here is that only the strictly congruent MNs would actually successfully match an observed action with the activity patterns that are present when the same action is executed. If understanding an action involves (or, more robustly, just means) being in state that one is in when performing the action, then MNs are a highly unreliable means of understanding.

Response to criticism 3

This is a tricky issue, since Gallese and other proponents of a robust role for MNs in social understanding interpretation point to the same data in support of their position. Gallese, for instance, argues that broad congruence provides the independence from specific motor movements that is essential to a level of understanding abstract enough to qualify as action understanding. Speaking of broadly congruent MNs, he asserts that they are "especially interesting, because they appear to generalize across different ways of achieving the same goal, thus enabling perhaps a more abstract type of action coding." (Gallese 2001)

This is not a bad idea, since it is based upon the insight that understanding an action involves an appeal to something more general than the individual action itself. Traditionally, this something more general would be a category of actions under which the individual action is subsumed. At the very least, one should be able to bring knowledge of actions one has witnessed or performed in the past to bear upon the present situation. Typically, this would involve appealing to a class of action, which would involve abstracting from the details of the current action. Abstraction, in other words, is the key to generality, which, in turn, makes potentially useful information accessible. Gallese appears to acknowledge this in the passage cited above. On a more minimal construal, matching the action with a prototype of an action would enable one to reason about the action without abstracting from it; one would simply substitute a different, equally detailed, action representation for it. If the prototype action is fairly common, one could achieve generality this way.

In any case, just mimicking the movements or a certain subset of the neural activity of an agent is not in itself enough to achieve the generality needed for understanding. And it seems that broad congruence is indeed a step in this direction. How could this work? MNs sensitive to multiple similar actions could indeed constitute a neural realization of categorization, insofar as they could group together various actions and treat them as the same, thus effectuating generalization via abstraction. Unfortunately, this does not seem to work, since the sets of actions and/or movements to which they are sensitive do not seem to constitute sets of similar actions, but, rather, sets of functionally related actions.

Nevertheless, sets of related actions offer a departure from exact mirroring of specific movements, and could also work as a basis for some more general understanding. The basic idea is that MN activity during observation of an action would represent a chain of actions likely to follow upon the action being observed. Iacobini speaks here of the activation of "logically related" MNs (Iacobini 2005, 533). This formulation is perhaps misleading insofar as Iacobini is not talking about logical relations in the sense of formal logic. What he has in mind are the kinds of connections among actions that simply make sense to us in light of our everyday experiences – e.g. after grasping food, one often places it in one's mouth. Activating such connections would enable one to predict the agent's behavior. Hence, MNs could be said to be involved in simulating the agent's ongoing action for the purpose of prediction.

It is a further question whether MNs in this model could be said to achieve or contribute to achieving action understanding or retrodiction of a prior intention. Jacob, for one, thinks not (Jacob 2008, 211). I will come back to this further below (6.4.2). For the time being, I just want to point out one distinction that should be helpful in settling this issue. I think it makes a difference which of the following two cases obtains:

(1) Among the MNs that are active during observation of action, there are some which are also active during performance of functionally related actions and also some which are active during performance of the same action.

(2) Among the MNs that are active during observation of action, there are only neurons that are also active during performance of functionally related actions, and none which are active during performance of the same action.

In the former is the case, it could be argued that the generalization required for understanding is achieved by the MNs. Presumably, it would make sense to frame this in terms of prototypes of actions which serve as proxies that allow inductive inferences about ensuing actions. If, on the other hand, the latter is the case, then the MNs could not be responsible for choosing which chain of logically related MNs to set in motion. That would have to be achieved elsewhere. The utility of the MNs would then be contingent upon the prior achievement of something akin to categorization of the observed action. I will come back to this issue later.

To sum up, if there were no relation between the actions to which broadly congruent MNs are sensitive during performance, on the one hand, and observation of the other, then Csibra's criticism would indeed be quite damaging. But, given that the connections among the actions associated with individual MNs or MN groups reflect common sequences of actions in the world, it seems that broad congruence could be a part of an account of how MNs provide a kind of understanding that is sufficiently independent of specific motor movements to qualify as action understanding. But the issue is far from resolved.

6.4.4 Criticism 4

This is essentially the point mentioned briefly in the introduction, namely that data from studies with apes cannot directly be brought to bear upon theories about humans. Pushing this line of criticism further, one could point out that there are indeed theoretical reasons against drawing analogies between MNs in apes and the mirror system humans. Specifically, the arguments would be that since apes are much worse than humans at imitation and social understanding, whatever mechanisms they employ are obviously not enough to account for imitaiton and social understanding in humans. Making matters worse, the species that is most commonly used in these studies, namely the macaque monkey, is especially ill-suited to the purpose of learning about social understanding, since macaques are even worse than most primate species at imitation[72].

Response to criticism 4

The only reasonable response is to concede that human action understanding is much more sophisticated, but to argue that it probably builds upon primate action understanding. We should therefore start simple and learn what we can about the basic mechanisms that are likely to be quite similar among primate species, including humans, and then add in other ingredients as needed to account for humans' more sophisticated abilities. This is indeed the strategy proposed by Goldman and Gallese (1998). They write:

> Our conjecture is only that MNs represent a primitive version, or possibly a precursor in phylogeny, of a simulation heuristic that might underlie mind-reading (498).

Obviously, this leaves important questions open. Just what additional resources come into play in humans? What role do MNs play in conjunction with these other resources? But we are not therefore reduced to mere hand-waving. An important step to take in addressing these questions is to improve the integration of behavioral and neurobiological studies comparing humans with non-human primates. If we look a bit more closely at the primatological research in the past ten years, we find that the picture is not as clear as I suggested in formulating this objection. To find, there is converging evidence that non-human primates are able to understand conspecifics' states of mind (epistemic states, attentional states and intentional actions) to some extent, but that they are not too terribly interested in doing so. They do so only to the extent that others' minds are immediately relevant to the production of concrete events in the environment that they themselves are interested in. Humans are therefore not

[72] Tomasello and Call 1997.

unique in having some kind of folk psychology or understanding of others' mentality, but in being interested in others' states of mind for the sake of sharing attention and other states of mind and cooperating in activities just for the intrinsic fun of it (Tomasello 2008).

Not only do these findings somewhat undermine the objection, they also suggest a battery of questions that should interest folks working on MNs. If apes are able to understand intentional actions but not so terribly motivated to do so, then one wonders whether emotion areas such as the amygdala are less closely linked to MN areas in apes than in humans. Aside from the simple anatomic aspect of this question, the functional connections could be investigated by checking whether the amygdala is perhaps less active in apes than in humans during action observation, or whether other signs of emotional arousal are more prominent or more readily evoked in humans in various test conditions involving action observation.

False belief studies could be set up where the test subject has a greater motivation if the experimenter doing the searching finds the hidden object. The reward could be either a private reward for herself, or a social reward, such as an opportunity to commiserate with (share attention with) the experimenter who failed to find the object.

If apes are better at understanding actions when they have an individual motivation for doing so, it would also be interesting to know whether humans' motivation to understand action differs in its physiological profile (where the activation is, what emotions are aroused, how they are expressed in heart rate, muscular activation, etc.) in cases where they have a personal interest in understanding others' actions can cases where their motivation is "merely" social or hinges upon attaining a shared attentional state. If there is a difference, then the link could be relevant to understanding how humans' use of MNs in action cognition differs from that of other primates. It is surprising that there has been so little cooperation between researchers working on these different approaches.

6.4.5 Criticism 5

Gallagher objects to the marriage of ST and MNs on terminological grounds that we have discussed (chapter 4) in the context of his critique of ST in general. The objection rests on an interpretation of the concept of simulation as being akin to pretense. This is not unreasonable, since the concept of simulation, as we have had ample occasion to observe, is much in need of precision, and starting out by looking up the definition in a dictionary is a fair move (more on this in another chapter). But it is also fair to depart from that definition as long as one is clear about doing so, and we will indeed be doing just that in the following. At any rate, terminology aside, Gallagher's objection raises important issues.

After criticizing on several grounds the notion that we explicitly simulate others (in the sense of pretense) in order to understand their behavior very often in everyday life, he considers the possibility that we implicitly simulate them. Against this idea, he points out that MN activity is exogenously caused and not deliberate, whereas pretense, and therefore also simulation, is endogenously caused and deliberate. Gallagher therefore denies that it makes sense to think of MNs as actively simulating, pretending or being involved in imaginative processes. He thinks it is more reasonable to say that MN research reveals that motor activity turns out to be involved in perception, which is surprising and important, but has nothing to do with simulation. (Gallagher 2006)

Response to criticism 5

One is certainly tempted to say "So what?" and leave it at that. What difference does it make whether we call MN activity simulation or not if (this is a genuine "if") it accords with the basic idea of simulation theory. At any rate, we have plenty of options to respond to the terminological challenge. We could, for example, either alter the relevant concept of simulation to include MNs or make them into non-simulative components of simulation processes. But the criticism, in my view, does at the very least highlight one important issue. Pretense or deliberate simulation seeks to represent or model a separate system. And this is an aspect of the concept of simulation that we surely want to retain if we want to account for action understanding. Just happening to be in the same state or a similar state as someone else does not count as understanding that they are in that state. If MNs are automatically activated during certain perceptual processes, it is not clear that they refer in an appropriate way to the other person. This point is a genuine challenge to anyone who wants to defend ST with the help of MNs, and I will come back to it later on.

6.4.6 Criticism 6

Goldman (2005) attacks Gallese on just this point. He doubts that mirroring or resonance in itself qualifies as understanding. He argues that we need stricter criteria for saying when the mental state or event that arises in the observer can be regarded as constituting understanding of an agent's action. Goldman acknowledges that this representation could be the same state as the agent is experiencing, but that this matching is not enough to make it into a representation of the other's mental state, i.e. to be *imputed* or ascribed to the other (thus making it a meta-representation). Another way of putting this is to point out that understanding or representing is an asymmetric relation, whereas pure matching is not. (Goldman 2006, 37) The classical account, according to Goldman, would be that the agent's intention is subsumed under a concept, which has a general definition. If Gallese wants to avoid this, he has to give an alternative account of ascription. Goldman sketches out two criteria and challenges Gallese to fulfill them. Goldman writes:

> Two people each undergo a bout of nausea while sailing on an ocean vessel, one in the Atlantic and one in the Pacific. The two are not in communication; they don't even know of one another's existence. Does either person understand the other's mind? No. (Goldman 2005)

Thus, Goldman makes two demands. *First*, the representation in the observer should be caused by the representation in the agent. This is Goldman's first stab at making the representation have the other person's representation as its content (it has to somehow refer to the other person). Goldman, interestingly, initially made a name for himself by espousing an externalist epistemology according to which knowledge does not have to be justified true belief, but only true belief that is causally brought about by the state of affairs believed to hold. There is no need to go into this at great length here, since obviously there is no reason to think or to demand that our understanding of others in folk psychology always or even often satisfies strict standards of knowledge. Furthermore, there is no reason why he should be wedded to this way of making a representation have someone else's intention as its content; this was simply an obvious option that occurred to Goldman naturally since it is apparently his favorite way in general to get content. In other words, he has recycled an old pet idea. Fine; let him.

Goldman's *second* demand is a bit trickier – namely that the representation should have an evolutionary *function* that depends upon its being reliably caused by the other person's mental state. He needs to add this in order to rule out cases of mental contagion – i.e. someone else is jolly and by talking to him I, too, become jolly. If mental contagion is not a case of understanding, then the causal connection is a necessary but not a sufficient condition. If, however, I have a mechanism that evolved because it causes me to enter into the same state as an agent I am observing, then presumably this mechanism will have evolved because it makes a difference of some kind for my behavior. Specifically, it is probably connected with subsequent processes that regulate my behavior in a way that accounts for the other person's action. The mechanism is therefore instrumental for a well-adapted response to the other person, and can therefore be said to have the other person's action as its content.

As far as I can tell, Goldman regards this second demand as a further necessary but insufficient condition for understanding. Taken together, the two conditions would also be sufficient. It is certainly a helpful contribution that Goldman makes in giving us a set of precise criteria for action understanding, and he is justified in demanding that the issue of ascription be addressed by anyone attempting to account for the role of MNs in action understanding.

Response to objection 6:

Since Goldman and Gallese (and just about everyone else) differ on how to account for ascription, I would like to postpone the response to this objection until the sections in which I deal with the positions presently espoused by Goldman (section 6.5.1) and Gallese (section 6.6.2).

6.5 Subsequent theoretical advances 1: deflationary approaches

There are numerous ways of responding to the various points of criticism while retaining a role for MNs in social cognition. I will divide these options into two types. In 6.5, I will discuss the first of these types, namely versions in which the role of MNs is dependent upon a ascription of the agent's prior intention (i.e. a metarepresentation thereof) occurring elsewhere in the brain. In 6.6, I will discuss the other type, which consists in the more radical option of doing without a distinct metarepresentaiton to account for ascription. Goldman's recent publications exemplify the first type, whereas Gallese's post-1999 publications place him squarely in the second type. An important question for us to bear in mind will be the extent to which these theoretical alternatives are compatible with ST, and if so, with what version of ST.

6.5.1 Goldman: A role for MNs in conjunction with mental concepts used to ascribe prior intentions

Goldman's (2006) basic idea is that MNs assist in understanding actions, but that other perceptual information also comes into play, and that action understanding involves ascription of an intention, i.e. application of a mental concept. The role that MNs play in retrodicting prior intentions (high-level mindreading) remains as in 1999: A hypothesis is formed about what intention the other person may have, and then a simulation begins, whereby an appropriate action is chosen with the help of MNs. If this action matches what the other person is doing, the hypothesis is confirmed and action understanding is achieved. Goldman does however allow that MNs or other mirroring mechanisms may suffice for understanding some basic actions (low-level mindreading).

It is apparent, then, that Goldman adds not only contextual information but also knowledge of the agent's mind, i.e. mental concepts. It is important to recall, however, that this account yields a highly unusual construal of mental concepts. According to the standard functionalist (or theory theory) construal, mental concepts are defined by the psychological laws in which they occur, i.e. their nomological relations among each other as well as to perceptual inputs and to behavior. But for Goldman, using one's own decision-making procedure (in some cases MNs) to get from a goal representation to a movement obviates the need to represent the psychological laws linking the agent's intentions to her actions. Hence, mental concepts must be defined somehow independently of their connections to behavior. Goldman winds up with an introspective account of mental concepts, which is certainly a minority position.

6.4.2 MNs in prediction: Jacob's proposal

Although Jacob and Csibra are among the sharpest critics of Goldman and Gallese (1998), both attribute an important role to MNs in social cognition. Jacob starts out by arguing that since motor intentions do not stand in a one-to-one relation to prior intentions, retrodiction of a prior intention can be achieved only with the help of activity in others areas beyond the motor system, notably in the superior temporal sulcus (STS) Cells in the STS lack motor properties but do have perceptual properties in common with MNs, i.e. they fire during the observation of actions performed by conspecifics (Jacob 2008, 193; Perrett 1989; Keysers and Perrett, 2004). Beyond this, they also fire in response to head- and eye-movements of conspecifics. Given that head- and eye-movements are obviously especially relevant for tracking others' attentional states, these cells could well play a role in detecting prior intentions. Jacob therefore refers approvingly to a paper by Allison et al (2000) in which it is claimed that "there may be a purely perceptual network of 'social perception' involving the STS, the amygdala and the orbito-frontal cortex" (Jacob 2008, 217). Jacob regards the fact that MNs do not seem to respond to head- and eye-movements as support for the view that MNs could compute the motor commands and therefore the movements that would best serve to carry out a prior intention *once a prior intention has been ascribed*. Thus, Jacob grants the MN system a prominent role in predicting motor intentions and thus also behavior, but makes their contribution dependent upon the representation of a prior intention, which is not achieved by the MNs.

It is important to note that he is in agreement with Goldman that prediction has to proceed from ascription of a prior intention. So is Jacob's proposal compatible with Goldman's version of ST? Jacob himself does not seem to think so. He interprets Goldman's theory as asserting that MNs are involved only in retrodiction and not prediction, and therefore takes himself to be disagreeing with Goldman, since his own view is that MNs are only involved in prediction. He writes:

> …either MN activity is predictive (in accordance with the new model of logically related MNs) or it is retrodictive (in accordance with Gallese and Goldman's conjecture). In accordance with the new model, it is, I think, more plausible to choose the former option, on which MN activity is predictive (Jacob 2008, 213).

I think this remark betrays a slight misconstrual of Goldman's position. It is true that for Goldman *the perceptual properties* of MNs serve only in retrodiction. But he also thinks that the same neurons can be involved in prediction. Neurons in F5 link intentions to movements. When the movement is observed, the intention can be inferred (in high-level mindreading, Goldman agrees, more information must be involved as well). When the intention is represented, movements can be predicted. Goldman talks more about retrodiction not because he restricts MN activity to retrodiction, but because retrodictive function is what distinguishes MNs from other neurons in F5. All motor and pre-motor neurons should be able to carry out the predictive function. So, from Goldman's perspective, this is not what is interesting about MNs and is not worth going on at great length about.

So, given that Goldman distinguishes between high-level and low-level mindreading, and agrees that high-level mindreading involves more than just the activity of motor and pre-motor systems, the differences between Jacob and Goldman appear minor. Jacob's proposal turns out to be compatible with Goldman's view that MNs, like other motor and pre-motor neurons, are involved in predicting action on the basis of the intention. Indeed, Jacob's view could even be made to fit with Goldman's view that MNs are *involved* in retrodiction, although he himself is skeptical of the putative role of MNs in retrodiction. I see two ways of making this work, which are compatible with each other, but each of which could be true independently of the other.

Firstly, the retrodiction of the other person's intention could make use of perceptual information about the situation as well as perceptual information about the other person's movements, the latter of which would be achieved with the help of MNs. The retrodiction would not take place in the pre-motor cortex, but would make use of information deriving from there, as well as information from other areas. In fact, information from all kinds of other sources could of course also be accessed – short-term and long-term memory, knowledge about the other person, language processing, etc.

Secondly, there is no reason why MNs should not have the predictive function Jacob ascribes to them, and *additionally* serve to test conjectural retrodictions in just the way Goldman thinks they do. It is highly plausible that these two functions could be related. Indeed, MNs would be doing exactly the same thing in both types of case, namely computing movements to realize prior intentions that are represented elsewhere, just as Jacob says. The sequence would therefore be:

1. Perception of the other person's movements and of the situation (leaving open what role MNs may play in this)
2. Conjectural retrodiction of an intention
3. MN activity derives movements to realize that intention
4. Comparison of the movement predicted by the MNs with the movements observed (leaving open what role MNs may play in this)
5. If there is a match, conclude that the conjectural retrodiction was right; if there is no match, then return to step 2 and try retrodicting a different intention.

I think it is a telling fact that Jacob overestimates the difference between his own position and that of Goldman. This overestimation is based upon an unnecessarily sharp distinction between prediction and retrodiction. Goldman deliberately does not make such a sharp distinction; for him the very same MN activity could be used in prediction and retrodiction. By making such a sharp distinction and limiting MN activity to prediction, I would argue, Jacob implicitly denies that MNs have perceptual properties at all. If we assume for the moment that Jacob's explanation of the function of MNs is right – namely, that they compute motor commands from representations of prior intentions, then what they are doing does not really have to do with perception at all. It is true that they would often be active while a subject is observing someone else, but this would not be because they have perceptual properties but, rather, because one is often predicting others' future behavior while one is observing them.

This point comes out in Jacob's denial of what he calls "strong congruence between the motor and the perceptual properties of MNs" (211) What he is denying here is not exactly the same as congruence in the sense in which I have already discussed it, i.e. as an *empirical* measure of how closely the perceptual and the motor properties of particular MNs or MN groups match. Rather, his claim is of a conceptual nature. To whit, he thinks that since the activation of MNs has a predictive function, MNs do not resonate with the other person's movements; rather their activity should correspond with likely upcoming movements of the other person. Thus, congruence, as he puts it (I would prefer to say resonance), would not even be helpful. He is therefore denying that there is any congruence (resonance) at all. I would add that this amounts to an implicit denial that it makes sense to speak of perceptual properties of MNs. Since MNs should, in Jacob's view, only be involved on the prediction side, it would be a mere coincidence that they should also be active during perception. Hence, his model implicitly denies the standard interpretation of MNs as having perceptual properties at all.

The trouble is that it is not immediately clear how to test between these two alternatives, since Jacob's model would in fact predict that MNs would also (coincidentally) be active during perception. Indeed, since one's predictions about people's future behavior would tend to be continuous with one's observations of their current behavior, Jacob's model even predicts that the motor neurons that are active (coincidentally) during perception of others' behavior should pretty closely match the motor neurons that would be active while one performed the action being observed. It seems, though, that the greater the congruence, the more plausible is the interpretation according to which they do in fact have perceptual properties.

Theoretically, it seems questionable to conceive of representation and/or ascription of a prior intention as being independent of a motor plan for realizing that intention. Aside from this, the *empirical work* on mu rhythms discussed in section 6.2.2 bodes ill for Jacob's account. It suggests that MNs play a role in action understanding that cannot be limited to the calculation of bodily movements for executing prior intentions.

6.6 Subsequent theoretical developments 2: mirroring as understanding without adding on separate mental concepts

There are also theorists whose interpretation is more robust that that of Goldman and Jacob in that they claim that mirroring does not merely play a role in but actually constitutes action understanding, i.e. without having to add separate mental concepts to the picture. Pineda and Gallese will be our primary examples of this approach.

6.6.1 "The extended mirror system"

Jamie Pineda presents (2008) a theoretical framework which situation mirror neurons within a broader range of mirroring phenomena in which a mirror neuron system (MNS) features as part of an "extended mirror system" that also includes other areas, such as the superior motor sulcus (STS) and the sensory-motor cortex (Pineda 2008). Pineda distinguishes a core MNS and an extended MNS. The core MNS consists, in his view, of the ventral premotor area (PMv) of the inferior frontal gyrus (F5 in monkeys BA 44 in humans), the parietal frontal (PF) in the rostral cortical convexity of the inferior parietal lobule (IPL).

The extended MNS includes other areas in which there are no mirror neurons, but which are anatomically and functionally linked to the core MNS. He notes that these other areas are essential for the "subsequent elaboration of the information" and that "the level of transformation performed on the data would make them critical to the outcome and part of an extended mirroring processes." (Pineda 2008, 4) One area that he considers to be especially important in this context is the superior temporal sulcus, (STS), which is reciprocally linked with the parietal frontal area in the inferior parietal lobule (which is in included in the core MNS system). The STS contains neurons that respond to biological motion (eyes, head and body), which is obviously relevant for identifying and interpreting actions. His inclusion of the STS is peculiar, since, as he himself notes, there are no mirror neurons in the STS. It may even at first glance seem to undermine the claim that MNs are important in action understanding that the STS play a crucial role in action understanding but ahs not MNs. But this would not be a fair conclusion to reach. It would be unreasonable to expect there to be MNs associated with interpreting eye and head movements. If you are following somebody else's eye and head movements, which is a sensible thing to do if you want to figure out what they are up to, then your eyes and your head will not mirror theirs, since you are not standing in the same spot as them. Hence mirroring would not be an efficacious strategy in this regard as it would in other regards. Therefore it should be no surprise that there are no MNs in the STS, and this fact should not undermine our confidence in the view that MNs contribute to action understanding when mirroring is an efficacious strategy.

I agree with Pineda that the involvement of these other areas does not exclude an active role for MNs even in cases where MNs are not the sole or primary contributor to action understanding. But I take issue with the claim that MNs plus these other areas constitute an extended mirror system. The term "mirroring" implies that action understanding is achieved when the observer is in the same state as the agent. But the empirical data is equally compatible with the view that there is an overlap in the states of the observer and the agent (namely MNs), but that understanding is achieved not by mirroring, but by categorization on the basis of MNs as well as other processes in the brain.

6.6.2 Gallese

The other somewhat radical way to retain a role for MNs in social cognition is to try to show that MNs indeed suffice for action understanding – at least for one kind of action understanding – without adding on separate conceptual resources. Gallese toys with this option. He spells out his position by postulating a special kind of understanding, which he calls "experiential understanding". What distinguishes experiential understanding from conceptual understanding is that we model, or simulate, others' behavior and/or internal states. Gallese proposes that mirror neurons are the likely neural correlate of this kind of simulation. It is unclear how we are supposed to interpret the relationship between experiential and conceptual understanding. It could take the former as an alternative to conceptual understanding, or as a part of conceptual understanding. Gallese seems to acknowledge both possibilities, but focuses on simpler cases in which experiential understanding could be present in a pure form. It appears that he wants to explain as much of social cognition as possible in this way, and then add on conceptual resources as necessary. In any case, obviously, this is a version of simulation theory.

Although I am calling this a radical alternative, it is more accurate to distinguish two versions of this position, one of which is much more radical than the other. The difference hinges upon the answer one gives to the question whether to regard "experiential understanding" as constituting concepts in a unique way or as non-conceptual. The latter is obviously more radical. As I mentioned, Gallese appears to toy with this latter option, but wavers. In some places, he asserts that the motor system obviates the need for mental concepts or for a representation of the other's intention, while in some places he apparently asserts that activation of MNs itself constitutes the use of mental concepts or representations of intentions. As for the more anti-conceptual interpretation, Gallese, Keysers and Rizzolatti (2004) claim that "the fundamental mechanism that allows us a direct experiential grasp of the mind of others is not conceptual reasoning but direct simulation of the observed events through the mirror mechanism" (396) On the other hand, Gallese 2003 writes: "it is the natural function of natural information to produce intentional representations, *concepts included* (2003, 1232; original emphasis deleted, new emphasis added).

This is obviously borderline inconsistent. I think that Gallese really has the less radical version in mind, but slips into more radical formulations at times simply because he himself (and in fairness, the world at large) is not so clear about what concepts are and what role they play in action understanding. In any case, I will focus on the less radical interpretation not only because I think it is more congenial to Gallese but also because I think it is more promising. We will see later on, when we discuss simulationist theories of concepts (chapter 7), that it is possible to view concepts as being constituted by motor and modality-specific perceptual representations as opposed to amodal, symbolic representations. When one employs concepts, these motor and perceptual representations[73] are activated, thereby simulating actions and/or perceptual experiences. That is why I call them simulationist theories.[74] Importantly the term simulation here is expanded to refer not only to simulations of others' actions but also to one's own past or future actions and experiences. I promise that this will be spelled out and assessed in chapter 7. For the moment, suffice it to note that Gallese thinks of concepts this way. Although that may be a slightly unorthodox view of concepts (in philosophy anyway, although in psychology it is accepted as one of the leading conceptions), it nevertheless makes Gallese's position appear more conventional on the whole, since at least he is including concepts one way or another in his picture of action understanding. So, given this background, let me sketch the strategy that Gallese pursues.

What needs to be represented is the agents' prior intention, in other words, the goal of an action. In Gallese's view, an action representation is constituted by a combination of motor and perceptual representations, and not by any additional symbolic, amodal representation. Representations of intentions are not just MN activity, nor are they additional representations alongside the motor representation and whatever perceptual representations may be relevant, but a unique combination of them. One feature of this proposal is that the contents of the various representations that constitute the representation of a prior intention are states of affairs, processes and events (including those performed with one's own body) in the world, not in the agent's own mind.

But somehow minds must come into it. How else could we introduce the element of ascription that appears to be necessary for understanding others' actions, that is, to differentiate between understanding what one is doing (or imagining oneself doing) and understanding what someone else is doing? The observer may therefore represent the agent's prior intention by mirroring their motor system and also their perceptual states (as would anyway be the case if she is in the same situation as them, looking at the same objects, etc.), but then this constellation of representations still has to be ascribed to the other person somehow.

Gallese's basic idea is that that the MN activity issues in different subsequent causal effects in the first-person versus the third-person case. Gallese writes in response to Goldman:

[73] As well as proprioceptive information and various other resources. More on this in chapter 8.

[74] Although theorists in question do not call their theories simulation theories, they do all use the term "simulation" prominently and pretty uniformly.

> The activation of MNs in the observer's brain not coupled with the activation of other sensory-motor circuits normally activated during the execution of the same action, is sufficient to automatically attribute the action plan to another creature. Hence the imputation process could be a default feature of the non overlapping patterns of brain activation during observation vs. execution of the same action. This of course does not deny the relevance of willful and conscious imputation driven by mental pretense and imagination, what Goldman qualifies as "second category of mental simulation mechanisms". It simply implies that there might be cases when this conscious process is not required. This is what "direct experiential understanding" is all about
> (Gallese 2005).

The idea seems to be that the representation of the other person is simply juxtaposed to the goal representation and one accordingly expects the other person to perform the action. This happens automatically, unless other sensory-motor processes are active that are active when one is performing an action oneself. This answer leaves some open questions, of course. Just what kinds of sensory-motor activation is specifically absent in third-person cases? Why is it absent? How does the automatic imputation occur? What kind of representation of the other person has to be present? Gallese does not give thorough responses to these questions. But the basic answer does seem to satisfy Goldman's demands. Presumably, the setup Gallese describes works because it was shaped by natural selection. The differential linkage of MN activity with subsequent brain processes just has to distinguish between one's own and another person's actions well enough to enable one to behave appropriately, and Gallese's account is at least intended to satisfy this, even if Gallese does not spell this point out in much detail.

Gallese's proposal also yields one particularly strange result. Saying that the ascription of an action representation to someone else occurs by "default" in the absence of a specific signal that one is performing the action oneself suggests that ascription is the *primary* function of MN activity, and self-ascription is parasitic upon or at least somehow ancillary with respect to third-person ascription. This is certainly a surprising thing to say about neurons in the pre-motor cortex. Which does not make it false, but it should make us think twice about it. To clear up one possible misunderstanding: Gallese cannot have anywhere near the same thing in mind as a functionalist who says that self-ascription is on a par with third-person ascription. For the functionalist, mental states are defined by their nomological (functional) relations, so one's grasp of one's own mental states is also mediated by one's behavior and one's perceptual inputs just as third-person ascription is mediated by observation of others' behavior and their perceptual input[75]. But this is of course just what Gallese wants to avoid with his notion of direct experiential understanding. He has to mean that we learn the meanings of the states in question in first-person experience, but that these states just happen to wind up taking on the primary function of third-person ascription. Certainly, this idea needs to be spelled out more clearly.

Interestingly, this is in a certain sense the converse of the sophisticated kind of functionalism espoused by Sellars. A functionalist could say, as Sellars does, that the nomological relations fix the meanings of mental terms, but that we learn to correlate qualia or something of the sort with functional states and can then self-ascribe on the basis of qualia.[76] While Gallese says that we learn the meanings of mental states by direct first-person experience and then learn to use these same states to ascribe mental states to others, Sellars presents a picture in which we learn to ascribe mental states to others and to oneself by other means, and then in a second step learn to self-ascribe on the basis of direct experience.

Maybe these accounts can be made to fit together in a way that combines the virtues of functionalism and theory theory with those of simulation theory. This would of course amount to the holy grail of the folk psychology debate, and so we ought not to get our hopes up too high, but we should bear it in mind as a theoretical desideratum as we attempt to refine our conception of the role of MNs in action understanding.

6.7 Some remaining concerns

There is some question about whether simulation theories of concepts are well-placed to explain the systematicity of thought. Simulation theories of *mental* concepts may also have problems whenever more complex reflection about the agent's states of mind would be necessary (e.g. false beliefs, incompatible desires). Another limitation is that it seems only to work when the goal state can be identified perceptually. Let me say a bit about this with respect to Gallese's interpretation of MNs.

6.7.1 Systematicity: Evans' generality constraint

I would like to conclude the section on Gallese by pointing out a different objection to Gallese (which also applies to Gordon's ST), which may be more serious. It is based upon a clever idea that Gareth Evans (1981) came up with. Evans is talking about when it makes

[75] This is sort of a caricature of functionalism and does not do justice to all functionalist positions: Sellars, for example, has a more sophisticated account. I will come back to this below.

[76] Sellars 1956, see chapter 2.

sense to ascribe abstract knowledge to someone. Specifically, Evans considers a language with twenty subjects and twenty predicates. We could ascribe to someone who masters this language either knowledge of all 400 sentences, or knowledge of the combinatorial rules. On the face of it, it seems that there is really no empirical way to test between the two hypotheses. Evans argues however that there is; we simply need to observe how people learn the language. If we observe that once they have learned to use a term (e.g. a subject word) in one or two sentences, they suddenly are able to use it in combination with all the other terms they know (in this case, with all the predicates), then we should ascribe to them abstract grammatical knowledge.

I would suggest that one (i.e. Goldman, but also TT defenders) could run the same argument against Gallese (and also against Gordon's ST). If people learn new actions and then suddenly are able to predict those actions on the basis of diverse perceptual information, in various situations, and with various specific motor realizations, there are good grounds for ascribing to them some single, unified representation – i.e. a concept of that action that is distinct from all the specific motor and perceptual neural activity that would be involved in these situations. I am not sure how to test this out, or how it would turn out if it were tested, but it does seem to point to a weakness in Gallese's position. Gallese is committed to saying that for every instance of understanding, there is a specific combination of motor and perceptual activity. The danger is that the possibility of generalizing may thereby be obviated.

6.7.2 Abstract goals

Recalling our assessment of embodied theories of concepts, we should point out that mental concepts often have just that feature that we said give his theory the most trouble, namely an independence from perception, or reference to non-perceptual information in their definition. Just think of the intention to insult someone. The goal is that the person feel insulted. Perceptual information may sometimes tell us whether this has been achieved (they cry or turn red, etc.) but often it will not. If perceptual information cannot distinguish between cases when the agent's intention is fulfilled and cases when it is not fulfilled, simulationist theories do not seem to be in a position to help Gallese.

6.7.3 Action on the basis of false beliefs

Another kind of case in which Gallese faces difficulties is when the agent is acting upon false beliefs. Here, the goal state can be identified perceptually, but the perceptual information that gets juxtaposed with the movement and with the other person has to be de-coupled from one's own perceptual information in order make the correct prediction. Given a classical false-belief task as our scenario, the observer would have to use the perceptual information that he had before the object was transferred in order to correctly predict where the other person will search. If he does not make this prediction correctly, he will not understand the agent's currently observable behavior as being directed toward a (hidden) object. It will just be non-goal-directed movement. Hence, he will not ascribe any prior intention at all and will not understand it as constituting an action. It would be very interesting to measure the activity of MNs in false belief tests. Apes, for example, have yet to pass any version of the test. Presumably, their MNs would be inactive during these tests. MNs in human children under about 3.5 years, who generally fail, should also be inactive, whereas MNs in older children or adults should be active, if my analysis is correct.

6.8 How do other versions of ST fit in with MNs?

6.8.1 Gordon

Gordon himself does not seem to have taken all that much interest in MNs, which is a bit surprising, I think, since it seems to me Gallese's interpretation of MNs fits best with Gordon's version of ST.

Aside from what I have already said about how Gordon fits in with Gallese, there is one other aspect that Gordon himself brings up in connection with MNs (Gordon 2005), Gordon remarks that the discovery of MNs suggests that simulation could be a natural kind. This may be interpreted as a pretty vacuous claim, i.e. if it is taken only to mean that there are various processes that can be called simulations. The notion of natural kinds is itself quite controversial, so it is not clear that we gain in clarity by appealing to it. But it is worth thinking about for a moment.

Whether we sue the terminology "natural kinds" or not, we should group together various processes and call them simulations only if gain something by doing so. In the case of natural kinds, we gain by being able to draw inferences about some member of the category by virtue of its belonging to the category. In other words we can project properties of one member upon the others. This (certainly controversial) constraint could also be a useful constraint upon using the term simulation. If we call MN activity and also emotional contagion and various other high-level imaginative processes simulative, does it enable us to project (some of) their properties across the board?

If so, then the obvious next question is why that works? We could compare their functions, their neural substrates, their development, etc. and search for commonalities to explain their similarity. Embedding the concept of simulation in such a framework could help to establish it as a scientific term and dispel the suspicion that it is simply a catchy term that is used to refer to lots of different things that, taken together, do not constitute a coherent position or an alternative to functionalism. So although Gordon's suggestion, as it stands, is not too terribly useful, it could be used to provide a framework for tying together various empirical and theoretical endeavours. But this is work to be done, and Gordon does not appear to have gone about doing it.

6.8.2 Heal

It is perhaps no surprise that Jane Heal has not been engaged in the discussion about MNs. For one thing, as we saw in chapter 4, her version of ST focuses on rational thought, which is fairly distant from the grasping-behavior and other simple action that have been looked at in connection with MNs so far. For another thing, her argumentation has an aprioristic character, since she claims that understanding others' psychological states must *in principle* involve simulating those states. For this reason, empirical work cannot be relevant in the evaluation of ST anyway (Heal 2003). Although her version of ST has its merits, one can at any rate conclude that it does not profit from the discovery of MNs, which, as I have argued, provides significant empirical suupport for other versions of ST.

6.9 Conclusion

Motor activity underdetermines intentional action. Aside from MN activity, it seems that other resources must be involved in understanding others' actions. To whit, there are at least two other kinds of information we also might invoke: contextual knowledge and knowledge of the agent's minds, possibly involving mental concepts. Let me say a bit in conclusion about the different ways we can add these two elements and how they fit together with ST.

As for the former – contextual knowledge – there are a few ways to incorporate it. Some, such as Iacobini, seem intent on a maximally robust interpretation of MNs, according to which MNs would take up perceptual information from elsewhere and calculate an intention on the basis of their own motor mirroring activity and the perceptual information.

So, here, MNs are not alone sufficient to ascribe an intention, but given the information coming from perceptual areas, they do in fact perform the ascription.

Another possibility is the view I have attributed to Gallese, according to which the ascription of a mental state just is a certain constellation of activity in various areas, without the various strands having to be brought together and an inference drawn on the basis of them. On this view, mental concepts can be said to be involved, but only in the sense that a functional description of all this activity would involve reference to the observer's meta-representations of the agent's mental states, obviously using mental concepts. This view could be dubbed meta-functionalism. Just as functionalism claims that ascribing mental states is the best way to account for people's behavior and is justified because the mental concepts refer to functional roles in cognitive systems, here we would say that the best way to account for a special kind of behavior, namely for people's accounting for others' behavior (folk psychology), is by ascribing meta-states to them. These ascriptions are justified by certain functional roles that constellations of brain activity play in regulating their own behavior (in light of others' behavior).

Both of these options are compatible with Gordon's version of simulation theory. It does not really matter much to Gordon to what extent the brain activity involved in understanding the agent's action occurs in the pre-motor cortex. For him, there are two decisive points. The first point is common to all versions of ST, namely that the processes involved in understanding the agents' action are largely the same as we would use in deciding upon and planning our own action. That these processes include perception and often abstract thought is no problem at all. The second point is that mental concepts do not play a role – neither as theoretical entities defined on the basis of functional connections (as in TT), nor as objects of introspection defined by their qualitative feel (as in Goldman's ST).

Since, on his view, knowledge of other minds and mental concepts do not play a central role in understanding their actions, he would not agree with Jacob and Goldman that their usefulness depends upon a representation of the agent's intention. For this very reason, he owes us an explanation of what makes the simulation into an ascription, i.e. how it refers to the other person at all. Presumably, the best way to do this is the meta-functionalist strategy. The mirroring processes would simply be linked with other subsequent processes in third-person versus first-person cases, thus influencing by behavior in different ways. In short, it would have a different functional role. That may work just fine insofar as we are similar to others, but Gordon still owes us an explanation of how we account for others' divergence from us.

Then we have versions, such as Jacob's, which fair better with regard to this point. For them, mental concepts really are invoked somewhere in the brain in order to represent the agent's intention in addition to the activity in the MNs and in perceptual areas. These theories are compatible with Goldman's simulation theory. More complex actions may indeed involve psychological knowledge, and would therefore call for a hybrid account such as Goldman (2006) has recently proposed, but with respect to the kind of simple examples that are currently being studied in MN research, this can be left aside. The key simulationist idea that survives here is that the observer's decision-making and action-planning resources are used when he seeks to predict or understand the agent's behavior. The same goes for Jacob, in fact. In the case of Jacob, the observer's mirror system predicts the ongoing motor realization of the other person's prior intention on the basis of probabilistically linked chains of MNs. The probabilities, in this case, would presumably be a function of one's own past actions, i.e. dispositions to act.

I have argued against Jacob's interpretation on theoretical and empirical grounds. Theoretically, I maintain that he divorces action understanding too radically from prediction of bodily movement. Empirically, his account does not do justice to experimental data suggesting that MNs are directly involved in action understanding (Oberman et al 2005, Pineda et al 2008). As for Goldman, he winds up with an introspective account of mental concepts, which, in my view, needs to be spelled out in more detail.

At any rate, the activity of MNs in action observation lends support to ST. Although the interpretation of MNs is contentious and far from settled, they do constitute an empirical finding that ST would have predicted, whereas TT would have had no reason to predict heir existence. If simulation theorists are to capitalize upon this empirical support, I would suggest that they need to accept that it constrains the available theoretical positions in the sense that ST should be formulated in a way that is as compatible as possible with MNs. The most ambitious way to do so would be to stipulate that ST should be formulated such that simulations could make use of MNs as their neural substrate. It could, for example, be hypothesized, that even high-level simulations (i.e. deliberate imaginative simulations) employ MNs or some similar processes. One could then propose specific empirical explanations of how such simulations work.

A weaker compatibility-demand upon versions of ST would be to stipulate that the various processes that are called simulations should be similar enough to qualify as a natural kind. This is of course a very dicey issue, since the notion of natural kinds is a very dicey issue. But the substantive point to be made here is that various processes could qualify as simulations if grouping them together enables us to project properties of one exemplar of the kind onto the others – for example, if functional or developmental explanations of certain simulation-processes turned out to work for other simulation-processes.

As to which version of ST is best supported by MN research, then, this remains unsettled. The matter depends upon what we have to add to MNs in order to get action understanding. If we add on other perceptual information about the situation, Gordon's ST suffices. But we have already seen in earlier chapters, and we have seen here again, that specifically mental concepts seem to have to play a role at least sometimes. Firstly, in order for matching to count as understanding here has to be something akin to ascription. Secondly, because understanding sometimes crucially involves not matching, namely insofar as the other person differs from the simulator. If we go for a sumulationist theory of concepts, as Gallese does, we can get ascription in a minimal way, i.e. without additional representations, and our interpretation will be somewhere in between Gordon's and Goldman's views. But an important point to bear in mind is that we can take this path only if we are willing to expand the concept of simulation to denote not only simulation of a target person's perspective but also simulation of our own past perceptual and motor experiences in the sense of an embodied theory of concepts. In the next chapter, I will discuss this approach in greater detail and also pursue the question whether such an approach can account for understanding others when they differ from us. In other words, the issue will be whether a simulationist account of mental concepts can be made to work.

Chapter 7:
Expanding the Concept of Simulation I:
Simulationist Theories of Concepts

7.0 Introduction

In this chapter, I would like to relate both TT and ST to the broader theories of concepts that they fit into.I have three reasons for doing this. Firstly, both TT and ST are theories of (or crucially include theories of) *mental concepts*. An evaluation of them should consider the merits and drawbacks of the broader theories of *concepts in general* into which they fit. Secondly, as we saw in the last chapter, a crucial issue in articulating the connection between ST and MN research is the role of concepts in action understanding – especially, but not only, mental concepts. Thirdly, the emergence of and the empirical evidence for what I call simulationist theories of concepts constitute support for ST, insofar as they reveal it to be a fruitful idea that fits with and also helps to guide empirical research and theory-building in the cognitive sciences.

It is safe to say that proponents of theory theory are relatively unified in sharing a functionalist view of mental concepts. There are two related theories of concepts in general that theory theory fits into: one is actually called 'theory theory' (Keil, Carey) and the other 'informational atomism' (Fodor). The empiricist strand of theory theory indeed arose within the context of a theory of concepts that is also called theory theory. The nativist strand should be regarded as belonging to the informational atomist camp. Proponents of ST, as we have seen, are less unified in their understanding of the nature and the role of mental concepts in folk psychology. Goldman, for example, has an introspectivist account of mental concepts, whereas Gordon tries to avoid mental concepts altogether, or to give them as behavioristic a gloss as possible.

Given this fact, it may be surprising to hear that there are also theories of concepts that prominently invoke the term simulation. As it happens, simulation theories of concepts can, I think, profitably be drawn upon in order to make out a role for mental concepts in folk psychology that accords with simulation theory, although it does involve an expansion of the term "simulation", as we shall see.

But first, let me say a bit about concepts in general. A good starting point is to view concepts as the constituents of thoughts (Margolis and Laurence 2008). This is a fairly uncontroversial baseline view that we can build upon. Ascribing thoughts to people is one important way of explaining their behavior in folk psychology as well as in scientific psychology, since people's behavior is influenced by thought processes like planning, reasoning, problem solving, deciding, recalling, etc. Thus, thoughts and thinking are of special interest to (folk and scientific) psychology, and so too are concepts insofar as they are the constituents of thoughts.

Of course, a good psychological explanation of thought and its consituents, how they relate to perception and to behavior, etc., does not automatically have philosophical significance. Philosophers certainly retain the option to claim that the concepts they are interested in are abstract entities (e.g. Fregean senses). A philosopher could maintain that the instantiation of concepts in human brains does not concern her, and can say nothing about philosophical issues, such as the intentionality, comopsitionality or the public/intersubjective nature of concepts. She can say that she is not looking for a description of how these phenomena actually relate to each other empirically, but for the most precise and consistent account. A good way of highlighting the distinction between psychology and philosophy here is to point out that psychologists should have no problem with incorporating contradictions in their theory, if actual normal human thought involves contradictions. Philosophers, on the other hand, regard contradictions as grounds to revise a theory.

I do not share this view. Being committed to naturalism, I would say that a philosophical theory of concepts gains immeasurably in plausibility if it is at least consistent with a/the leading psychological theory. If it is *based upon* a psychological theory, so much the better. The other side of the coin is that the reverse also holds. A psychological theory of concepts would do well to at least leave room for explanations of philosophical aspects such as intentionality. Moreover, psychologists should be (and are) interested in philosophical analyses as a source of explananda, since philosophical issues tend also to have empirical pendants. An elegant, empirically-based theory of concepts should also account for the relationship between concepts and things in the world (intentional content), their intersubjectivity (stability across various individuals) and their compositionality (the way they *in fact* combine in normal human thought). Thus, I favor the approach of looking for a theory that works both for philosophy and for psychology. But I will try to separate philosophical and psychological aspects, or at least to be clear about when I am refraining from doing so and for what reasons. At this point in this monograph, my primary interest is in an empirical account of concepts, so I am looking first and foremost at psychological theories, and will give psychological argumentation priority. Looking ahead to later chapters, the overall strategy will be to build a philosophical analysis upon this psychological basis, but that is not the focus of this chapter.

For now, let me start the discussion of concepts by setting forth a list of desiderata that a theory of concepts should attempt to explain. Of course, it may be that no one theory can explain all of them, or that the list requires revision in light of new empirical data or theoretical insights. Nevertheless, a theory of concepts should be judged by how well it accounts for these desiderata[77].

7.1 Desiderata for a Theory of Concepts

7.1.1 Scope

There is a great variety of concepts, ranging from the sensory to abstract, encompassing concepts for observable states, concepts for theoretical entities, formally derived mathematical concepts, etc. A theory of concepts should be able to integrate not only one kind of concept but as much of the apparent variety as possible. Empiricist theories (such as the simulationist theories that will be our focus) are generally taken to have special difficulties with this desideratum, since they derive concepts from perceptions, and therefore have trouble with concepts for which perceptual properties are not decisive.

7.1.2 Content

This has two aspects: *intentional* and *cognitive* content. As for *intentional content*, concepts refer to things (or properties or processes and maybe other stuff, too (there is no need to prejudge this here). The class of things to which a concept refers makes up its intentional content. Philosophically, a theory of concepts should include an account of their reference. This is essential in order to be able to give a semantic interpretation of thoughts and to evaluate their truth or falsity, concepts being, as noted above, the constituents of thoughts. Psychologically, accounting for this intentional content is explanatorily important because people’s behavior is sensitive to the intentional contents of the concepts constituting their thoughts. We may be able to explain the behavior of simpler organisms by appealing to occurrent stimuli, but humans can think about things, such as frogs, that are not immediately present, for example in order to make plans involving those things, such as a plan to catch and eat a frog. Hence, we must ascribe thoughts to them which represent and are somehow, namely intentionally, linked to those things. A theory of concepts ought to be able to account for this link to concepts’ intentional contents.

But intentional content is insufficient to individuate concepts. We can have more than one concept for the same thing or class of things (think of Frege’s morning star/evening star),

[77] I am borrowing this list to a large extent from Jesse Prinz (2002)’s clear, insightful, and well-balanced presentation.

or, conversely, more than one class of reference for the same concept (think of Putnam's Twin Earth example, which shows how the same concept, namely "water", applied in different contexts, can refer to different chemical compounds, namely H2O or XZY). The essential point here is that the role that things or classes of things play in people's thoughts is underdetermined by the properties of those things or classes of things (especially when those things don't exist!). The way in which people think about them, what they know about them, etc.[78], is also important, and must therefore also be a worthy target of any theory of concepts. I will call this the *cognitive content* of concepts.

7.1.3 Acquisition

If a theory of concepts also explains how people acquire concepts, so much the better. If it is not even compatible with any reasonable account of acquisition, that must be considered a minus point. Of course, this is a prime candidate for a point at which a philosopher is likely to say "I don't care". Fodor does basically that, for example.

7.1.4 Categorization

This is a biggy. Categorization enables us to draw inferences about the properties of an object on the basis of the category to which it belongs. It encompasses category identification and category production. In the former, a person identifies the category to which an object belongs, expressing judgments such as "That is a canary" or judging the truth or falsity of predicative sentences such as "Canaries are birds". In the latter, a person ascribes properties to an object on the basis of its membership in a category.

For psychologists, categorization has been by far the most important desideratum to be explained by a theory of concepts. Specifically, psychologists expect a theory of concepts to explain certain empirical facts about the way in which people identify and produce categories. "Typicality effects" are a prominent example. People rate some examplars of a category as more typical than others (Rosch 1975). These exemplars tend also to be categorized more quickly in categorization tasks (Rosch 1978). Another important empirical finding is a bias in favor of an intermediate level of abstraction, constituted by so-called "basic-level categories". For example, people tend to categorize a dog as a dog more readily than as a rottweiler (subordinate level) or as living thing (superordinate level) (Berlin and Kay 1969, Rosch, Mervis, Gray, Johnson and Boyes-Braem 1976), and can also perform inferences more

[78] The "etc." stands for the whole vast web of discussions on modes of presentation, sense and reference, intension and extension that has been ongoing since Frege's time.

quickly involving basic-level categories. Of course, philosophers are not obliged to care about these psychological effects, but I do.

7.1.5 Compositionality

This, of course, is another very important desideratum. It is invoked to explain the fact that the ability to form certain thoughts appears to automatically imply the ability to form certain other thoughts. This fact has two sides: namely productivity and systematicity, which go hand in hand. Productivity is the feature of concepts by virtue of which they can be used in forming new thoughts. Systematicity is the feature of concepts by virtue of which they retain their (intentional and cognitive) contents when combined with other concepts to form compound concepts. Thus, the contents of compound concepts are functions of the contents of their constituent concepts. Fodor (1994, 1998) is especially insistent upon the importance of compositionality. So, if you can form the thought, "Fodor prank-called Dennett", you should also be able to form the thought "Dennett prank-called Fodor", since the constituent concepts retain their contents as well as their syntactic features across various contexts. Systematicity is important to philosophers because without it a conceptual system may not be consistent. Psychologists are interested in it because it is an empirical fact that thoughts have compositional structure at least to some extent.

7.1.6 Publicity

Concepts should be public in the sense that they must be at least approximately shared by members of a linguistic community. Otherwise, communication would be impossible. Moreover, they could not otherwise be used in giving intentional explanations of behavior. A typical intentional explanation might be "Mary opened the cupboard because she desired a glass of scotch and believed there to be scotch in the cupboard." In order to find this explanation satisfying, you must roughly share my concepts of cupboard, scotch, belief and desire, and also must expect Mary to share those concepts as well.

7.2 Recent theories of concepts

I will now present a few leading theories in order to give an overview of the theoretical options. I will start out with imagism, since it is historically prior to the other theories I will present and also because it is the theory that simulationist approaches are reviving. For this same reason, I will go into a bit of extra detail in discussing its strengths and weaknesses.

7.2.1 Imagism

This is the classical empiricist position. In a nutshell, it asserts that concepts are derived from perceptual states. You encounter a tree, thus producing a perceptual state, which leaves a memory trace. Later on, when you think about trees, that memory trace is called up. Barsalou, one of the leading proponents of simulationism today, himself claims that his theory marks a return to perception-based theories of knowledge that dominated for 2,000 years until the twentieth century, and mentions the British Empiricists as exemplary of the view that cognition is imagistic. According to Barsalou, this approach came into disrepute through the efforts of behaviorists and ordinary language philosophers, who sought to banish mental states from scientific psychology. Barsalou explains this development:

> Because perceptual theories of mind had dominated mentalism to that point, attacks on mentalism often included a critique of images. The goal of these attacks was not to exclude images from mentalism, but to eliminate mentalism altogether. As a result, image-based theories of cognition disappeared with theories of cognition (Barsalou 1999, 578).

Although it is certainly legitimate to establish a link to earlier perceptual theories, Barsalou's historical sketch, I think, underestimates the novelty and the merit of the perceptual-based approach that he himself advocates (and which J. Prinz endorses). We will see why later on; for now, let it suffice to say that imagism is the historical precursor to Barsalou's theory.

Imagism has certain advantages. It automatically incorporates an account of *acquisition*, since it has concepts originating ontogenetically from perceptual states. As for phylogenetic acquisition, imagism is also in good shape since it can piggyback upon an account of the evolution of perceptual systems. An explanation of *cognitive content* also comes rather easily, since one may have two different images of the same object (i.e. the evening star shrouded in gathering darkness, the morning star emerging from a rosy sky at dawn) or have one image corresponding to various intentional contents (the same image of a clear, tasteless liquid for H2O and XYZ).

Imagism is also fairly well-placed to explain *categorization*. With respect to category identification, the task is to subsume a percept under a concept. If concepts are qualitatively similar to percepts, all one has to do is match a percept with concept most similar to it. With respect to category production, the task is to infer properties of an exemplar on the basis of its membership within a category. This can be achieved by simply reading off the properties of the image that constitutes the concept. So, if I know that a given exemplar is a dog, I can refer to my dog image to predict further features of the exemplar.

An explanation of *typicality effects* comes naturally to imagism, since percepts/exemplars that are more similar to my concept will seem more typical of that category. A hairless dog, for example, will be less similar to my dog image than a respectably furry dog, and will thus seem less typical. What about *basic-level categorization*? This is where the problems begin. Imagism seems to predict that categorization would be most efficient at a lower-level of abstraction. If concepts are mental images, then, like normal pictures, they should be unable to depict something as belonging to an intermediate level category without also revealing more specific information about the exemplar in question. Can one depict a dog without also depicting it as a certain kind of dog? Or depict a shape as a triangle without also depicting the relative sizes of its interior angles? And yet we are more efficient at identifying and producing categories at an intermediate level of abstraction.

This point harkens back to a problem with Locke that Berkeley (1710) already noticed. There is a tension between thinking of concepts as mental images, on the one hand, and thinking that they can abstract away from specific differences in order to represent an entire category (i.e. putting aside the differences between scalene, equilateral and isosceles triangles in order to represent something as a triangle). The link to specific perceptible exemplars also threatens the *scope* of imagism: how can abstract or theoretical concepts such as justice, the number 19 and antimatter be represented by images? A famous experiment conducted by Frank Keil (1989) also reveals the limitation in scope of imagism (although Keil's target was prototype theory, which is also impugned by the results of the experiment). Keil showed the participants a picture of an ordinary raccoon, and then had them paint a white stripe on it and add the kind of sac that a skunk has. After this transformation, the animal looks more like a skunk than a raccoon, but participants (from the fourth grade on) still insist it is a raccoon. Something other than their images of skunks and raccoons are obviously playing a role in their categorization here, presumably unobservable features such as innards, genetic constitution or phylogenetic background. These kinds of unobservable factors are difficult to account within the framework if imagism.

Publicity is also problematic insofar as different individuals may wind up having quite different concepts as a result of having experienced different objects. *Compositionality* presents another problem for imagism. There is no systematic way to construct complex images from simple images or to decompose complex images into simpler ones. So, for example, someone may have an adequate image of carnivores and an adequate image of plants but no adequate image of carnivorous plants. Conversely, someone can have an image of a carnivorous plant but no adequate image of other carnivores or other plants. This is closely linked to the problem of abstract concepts, since the problem is that that the decisive feature that makes one a carnivore is not something that can be perceived (at least not all the time.

Then there are some other classic philosophical problems that imagism would need to deal with in order to give an account of the intentionality of concepts. Think of the ambiguity of images. Wittgenstein (1953) is always a good source of this kind of example (the duck-rabbit, a man ascending a hill also resembles a man descending a hill). Some additional factor, such as an interpretive act, is apparently also needed in order to establish reference to the intentional content of an image. Moreover, even if resemblance were not undermined by ambiguity, it is not the same as reference, since it, unlike reference, is symmetric. Goodman (1976) is a recent locus classicus for this critique. A person resembles her portrait just as much the portrait resembles her, and yet it refers to her whereas she does to refer to it.

7.2.2 Definitionism

This is the view that a concept C consists of simpler concepts expressing necessary and sufficient conditions for falling under C (Margolis and Laurence 2008). Definitionism characterizes *acquisition* as a process by which simpler concepts are assembled to form complex ones. What about the primitive concepts that this account demands, though? Obviously, some other explanation has to be given for those. The stability of concepts across various combinations provides a good account for *compositionality*. In *categorization identification*, a complex concept is identified when an exemplar has the features required to satisfy the necessary and sufficient conditions. *Category production* would predict features of an exemplar once I know it belongs to a category.

On the negative side for definitionism, one would want to include procedures for predicting features that are often present but neither necessary nor sufficient, but definitionism appears to lack of way of building them in. Definitions can be thought of as senses, i.e. as the cognitive content of a concept. *Intentional content* is explained as the class of objects fulfilling the conditions. But if the conditions are specified in terms of further concepts (e.g. VIXEN = FEMALE, FOX), then we seem to be off on a hermeneutic boondoggle. At some point, at pain of solipsism, primitive concepts have to be linked to classes of things in the world. And no convincing account of how that might work has been forthcoming.

Definitionism has lots of trouble with *typicality effects* and has no explanation for *basic-level categorization*. Moreover, the philosophical daily grind makes it abundantly clear that definitions are elusive in most cases. Just think of Plato's ridiculous dialogues about justice and other such nonsense. Why should we think that we have definitions of all our concepts if teams of philosophers are unable to produce satisfying definitions of almost anything? People raising this criticism like to refer to Wittgenstein (1953), which also provides the point of departure for the next major theory of concepts:

7.2.3 Prototype theory

Prototype theory was designed (Rosch & Mervis 1975, Rosch 1978) to explain typicality effects. It posits that concepts are represented as best instances, or prototpyes of a category. So a concept C does not have definitional structure but probabilistic structure: a thing falls under C if it has a sufficient number of properties, which can also be weighted differently (Margolis 2006).

Compositionality presents problems for prototype theory, since, if it is to explain compositionality, it should predict that the prototype of a complex concept retains the features of the prototypes of its constituent concepts. But this often fails. To take Fodor's (Fodor & Lepore 1994) example, the pet fish prototype encodes features, such as being brightly colored, that are not present in the prototypical pet or the prototypical fish.

There is, on the other hand, some work that bodes well for a prototype-account of compositionality. Hampton (1988) found a class of cases that seem to violate compositionality but can be explained with prototype theory. Take the concept *tools that can be used as weapons*. Compositionality would appear to demand that something that belongs to this concept also belongs to the concept *tool* and the concept *weapon*. And yet people often diverge from this. Screwdrivers, for example, are consistently judged to be tools that are also weapons but not weapons. Hampton explains this by saying that people judge something to belong to the category if it has more than, say, half the requisite features of the prototype. Okay, so if the tool prototpye is composed of features a,b,c and the weapon prototype of features d,e,f, then compositionality demands that the tool that is also a weapon prototype be composed of a,b,c,d,e,f. A screwdriver may fulfill a,b,c,d, and thereby satisfy more than half the requirements for tool and for tool that is also a weapon, but not for weapon. I am not sure about this. The tool that is also a weapon concept is a highly artificial construction, and I suspect that people interpreted it as tool that can be used as a weapon, which does not imply weaponhood. Hence, their refusal to categorize it as a weapon would not even appear to violate the compositionality requirement, so there is nothing to explain. But there could be other such cases, and Hampton's explanation reveals the powerful explanatory resources of prototype theory.

Prototype theory has bigger problems with intentionality, since it cannot easily explain such effects as revealed in Keil's raccoon experiment described above. Moreover, it cannot account for the relations between properties of its exemplars, such as the relation between birds having wings and their flying. The same goes for hidden properties, such as familial relations. These criticisms were, historically, the motivation for turning to theory theory:

7.2.4 Theory theory

Concepts stand in relation to one another as the terms of a scientific theory (Carey 1985, Gopnik & Meltzoff 1996, Keil 1989). Gopnik and Meltzoff spell out their view of theories in terms of structure, function, and dynamics. Structurally, theories are systems of abstract entities and laws. Functionally, they yield predictions and explanations. Dynamic aspects include accumulation of evidence, reluctance to acknowledge theory violation, ad hoc adjustments and theory change. On this view, *categorization* proceeds much like scientific theorizing. I learn about an exemplar in various ways – by observing its outward appearance as well as its behavior and gathering contextual information, and subsume it under the concept that best explains all the data.

Since the theory theory was designed to deal with the fact that theoretical knowledge, notions of hidden essences and causal principles can override perceptual similarity in categorization tasks – as demonstrated, for example, in Keil (1989)'s raccoon experiment – it is no wonder that theory theory does well in accounting for such phenomena. But the case for primacy of hidden essences and such is not entirely clear. Hampton (1995) conducted a similar kind of study that yielded the opposite result. Participants were told about an animal born to horse parents but which is fed zebra food and comes to look and act like a zebra. Only a third of the participants said that the animal remained a horse. Although this result gives us reason to be careful about interpreting Keil's experiment, it is quite difficult to interpret. It is not only outward appearance but also something internal to the animal that might be at work in making people inclined to classify it as a zebra, since, after all, it was given zebra food. The participants could be making their judgment on the basis of an expectation that the animal's diet has transformed its "innards" and that it now has not only the outward appearance but also the hidden essence of a zebra.

Granting that theory theory probably scores some points in accounting for the role of hidden essences and such, it still faces some other problems. For one thing, what is the relation between theories and concepts? If concepts are to be characterized as mini-theories, then it would seem necessary to characterize theories without appealing to concepts (i.e. without their being constituted by concepts, among other things. But one of the central points about theories that Gopnik and Meltzoff want to import to concepts is that they contain theoretical terms. Now, what are these if not a kind of concept?

What other problems does theory theory encounter? *Intentional content* is a bit elusive. If the intentional content of a concept is to be fixed by people's beliefs about the objects falling under that concept, then we wind up with the kinds of problems that causal theories of reference[79] emphasized. In order to pick out all tigers, we need to have a theory of tigers that gets at their essence. How many people have such a theory? Even if we apply lax criteria and let the theory be a rather schematic appeal to internal constitution or phylogeny, this is not unproblematic. How do we individuate tiger's innards without calling them "tiger innards"? And what is unique about tigers' phylogeny? Presumably that they have "tiger parents". Obviously, this is all quite circular unless it is supplemented by information that most people lack. So, do most people lack a concept of tigers? And what if most people's theory about a category of objects is wrong, for example if most people think that whales are fish and not mammals? Then their concepts would turn out to not to refer and thus to have no intentional content. This cannot be right. Moreover, there is another problem with intentional content that arises from the potential regress, mentioned above, that results from saying that concepts are made up of mini-theories, which are made up of other concepts, etc. Specifically, one has to stop the regress somewhere and introduce primitive concepts. And then one needs to give an account of how they refer, and theory theory has yet to produce any theoretical options for doing so.

Cashing out concepts in terms of theories also brings in some difficulties in accounting for *publicity*. If a concept is constituted by a (mini-) theory, then semantic holism enters the picture, since a theory involves numerous beliefs, which in turn include concepts that include theories etc. Aside from the obvious danger of a regress, which I have already pointed out, this interconnectedness of concepts raises the possibility that two people may have different theories about a particular category of objects, since some of the beliefs in their semantic network for that category are bound to differ. Taking the concept ANIMAL as an example, Fr. O'Patrick might have a theory about animals that rules out their having souls, and a theory

[79] i.e. Putnam 1975; Kripke 1980

about humans that ascribes souls to them. Since I do not think humans have souls, I have no problem calling humans animals. So it seems like Fr. O'Patrick and I have different concepts of animals.

A theory theorist can of course respond that we do not need to suppose that everyone has identical concepts in order to communicate. It probably suffices if we have similar concepts. Smith et al. (1984) take this line, for example, arguing for a notion fo concept similarity defined by the number of shared features tow concepts have. Indeed, there is empirical work, which I discuss in section 7.3.2, showing that concept sharing across individuals and even within the same individual over time may be overrated (Barsalou 1989, 1993). But this strategy is not unproblematic, as Fodor and Lepore (1992) have argued. Specifically, Fodor and Lepore point out that the notion of concept similarity proposed by Smith et al. presupposes a notion of concept identity. In order to say that two concepts share a certain number of features, you need to say that they have *the same* features. And in order to say that those features are the same, you need concept identity, which is what we were trying to avoid by introducing concept similarity. Clearly, this strategy needs some refinement in order to work.

Prinz (2002) claims also that theory theory has problems with *compositionality*. I am not too convinced by his argument, but here is how it goes: my theory of Harvard graduates who are carpenters includes the belief or inference that they are nonmaterialistic, but this property is not contained in my theory of Harvard graduates or carpenters (Prinz 2002). Instead, I arrive at this inference by further theorizing that is not in any way included in or inherited from the constituent concepts. Prinz's gloss: theories do not combine easily. This seems a bit unfair to me. It is true that the belief bearing upon the composite concept cannot be deduced from the beliefs about the constituent concepts. But that is a wildly strict requirement. I will call it hardcore compositionality. I am not sure that any concept theory, except maybe definitionism, can meet the requirement of hardcore compositionality. But, luckily, I do not see any reason why they should have to. My beliefs about those constituent concepts do at least constrain and guide my theorizing about the composite concepts. This should be enough. In my view, the potential problem for theory theory with respect to compositionality could stem from holism. If concepts are defined by mini-theories, which contain concepts defined by mini-theories, etc., then there is a lot of information that could be relevant when combining concepts. What procedure functions to choose among the potentially relevant pieces of information and thereby reduce the complexity?

7.2.5 Informational atomism

The core of informational atomism is the view that the content of a concept (at least of a primitive concept) is determined by its position in a causal relation to something in the world. This starting point, borrowed from informational semantics (Dretske 1981, Millikan 1984), has the advantage of fulfilling the criterion of accounting for intentional content right off the bat. To see why, take an example, such as the concept COW. COW has cows as its content because it is cows and not umbrellas or steak-and-cheese sandwiches that cause COW to be activated. This means, though, that the content has nothing to do with the intrinsic features of the brain state or mental symbol that instantiates COW.

But in order to get at publicity and compositionality, Fodor insists on atomism (2000). To see why, consider an example invented by Dretske (1981), which illustrates a distinction between indicators and detectors. A machine has 26 lights corresponding to the 26 letters of the alphabet. Inside the machine, there are 26 letter templates, each attached to one light. When one of the templates matches an input, the corresponding light is turned out. In Dretske's terms, each light is an indicator of one letter, whereas each template is a detector. Indicators are not structured and their forms or relations to each other are arbitrary, whereas detectors are structured; their parts correspond to letter parts.

According to Fodor, (1991, 1998), concepts cannot be detectors, because we may use very different mechanisms to identify an exemplar, which would undermine publicity. I may identify dogs by their bark and you may identify them by their bite, but once we have tokened them, the same set of properties is represented. Our concepts are correlated with the same things in the world. This may not be true of detectors, so concepts cannot be detectors. Rather, they should be equated with indicators that are activated by the detectors.

What about compositionality? Fodor also denies that detection mechanisms could be compositional. An optimal mechanism for detecting red and an optimal mechanism for detecting hair may not combine to form an optimal mechanism for detecting red hair (Fodor 2000). Hence concepts are not detectors. Indicators, on the other hand, according to Fodor, can be compositional. He uses the analogy to language (hence his famous "language of thought") in order to clarify how this works. Concepts, like words in a language, can be combined with other concepts to form complex strings. When this occurs, they retain their semantic properties and also their syntactic or formal properties. The semantic content of a concept, given the informational semantics that the theory endorses, is whatever property reliably causes it to be tokened. This does not change when the concept is combined with another concept. In the complex concept RED HAIR, for example, RED retains its informational link to red and HAIR retains its link to hair. One can predict the semantics of RED HAIR on the basis of the semantics of RED and HAIR.

The notion that symbols in the "language of thought" have formal properties that are analogous to the syntactic properties of words in a natural language also points to a way of addressing the issue of *cognitive content* (or "sense" as opposed to reference) Briefly, the mental symbols MARK TWAIN and SAMUEL CLEMENS have the same intentional content, because the same thing in the world causes them to be tokened, namely the guy who wrote Huck Finn. Nevertheless, they are distinct concepts because they have *distinct formal properties*, and thus combine differently with other symbols.

It is not clear that this strategy works, though. Prinz (2002) points out that the analogy to natural language cannot get all the work done that Fodor needs in order to explain cognitive content, since senses do not map one-to-one with words. Referring to Kripke (1979), Prinz notes that senses can be more fine-grained or more coarse-grained than words. As for the former, consider an example: Sally does not know that Farrakhan the violinist and Farrakhan the leader of the Nation of Islam are the same guy, so it is possible for her to have two separate beliefs, both of which would be expressed with the sentence "Farrakhan likes to perform for an audience".[80] The one belief would be about a guy playing violin in a concert hall, whereas the other belief would be about a guy ranting and raving on a soap box. Conversely, senses can be more coarse-grained than words, as illustrated by translation. Otto has the belief expressed by the sentence "Austria is beautiful" and the belief expressed by the sentence "Österreich ist schön". Although the words with which these two beliefs are expressed differ, they arguably have the same sense, since they presumably can be subsumed under the same psychological laws and lead to similar behavior[81]. In short, words in different languages may be synonomous.

The theory admittedly cannot explain acquisition, and thus embraces nativism. Fodor's argument (1975) starts out from the distinction between primitive and complex concepts, the latter being composed of the former. BACHELOR, for example, is a complex concept, being composed of UNMARRIED and MALE. BACHELOR can be learned by hypothesis formation using its components, i.e. one forms the hypothesis that BACHELOR means UNMARRIED MALE and checks this hypothesis against the available evidence. But one cannot proceed in that way in learning primitive concepts, such as RED, since there are no components in terms of which one could formulate a hypothesis. Fodor concludes that primitive concepts cannot be learned; they must be innate. And, as it happens, Fodor thinks

[80] The example is from Kripke 1979

[81] The example is from Prinz 2002.

that most lexical concepts must be primitive and thus innate. The argument, as summarized by Prinz, goes like this:

> Lexical concepts cannot be decomposed into defining features because good definitions are frightfully scarce, and they cannot be decomposed into nondefining features, such as prototypes, because those do not combine compositionally. It follows that almost all lexical concepts are primitive. But if most lexical concepts are primitive and primitive concepts are innate, then most lexical concepts are innate (Prinz 2002, 95).

Of course it sounds preposterous to say that concepts like spatula, neutrino and Christianity are innate, but Fodor bites the bullet and endorses this conclusion (Fodor, 1975). In fairness, "innateness" can be taken in a pretty deflationary sense, namely to mean that we are born with the ability to detect spatulas, and the application of that ability causes spatula symbols to be generated, with which spatulas are subsequently indicated. Nevertheless, one may be inclined to regard this rabid nativism as detracting from the merits of informational atomism. Fodor himself has subsequently backed away from it, and now thinks there may another way to explain the acquisition of primitive concepts that avoids nativism. Specifically, atomic symbols can be generated and then brought under the nomological control of whatever they represent (1998, chap 6).

Fodor is now more concerned with a different aspect of concept acquisition, which is closely related to the issue of publicity of concepts. Imagine that we accept SPATULA as a primitive concept. If nativism is right, then the primitive concept is an innate symbol in the language of thought. If not, then it is still an atomic symbol with no inner structure, and its relation to spatulas is arbitrary. In either case, how is it that everyone's SPATULA symbol gets lined up with the same thing in the world outside, i.e. with spatulas, and not with giraffes or licorice? Fodor does not heave a good way of resolving this.

Fodor's theory may also wind up overdoing it with compositionality. Specifically, it fails to account for how compound exemplars fail to inherit many of the features of their constituents. For example, we can generate feature lists for RUBBER BUNNIES, TIN FOIL SOFAS and TEN-LEGGED YAKS (Hampton 1991). These feature lists reflect the features of the constituents, but also leave some the features out. So, a feature list associated with RUBBER BUNNIES is likely to contain features like having big ears but not being fluffy. In short, Fodor's work on compositionality is a major contribution, but it is probably not the last word.

7.3 simulationist theories of concepts

As I mentioned in the introduction (7.0), there are theories of concepts that can with good reason be labelled simulationist, and which in fact fit well with ST, and help to make simulatinoist interpretations of MNs plausible. We will see that the expansion of the term "simulation" to encompass certain views of concepts does indeed help to make the idea intelligible that MNs provide an alternative to a certain kind of "disembodied" conceptual reasoning, which operates upon amodal symbols. Insofar as empirical research may find that we use matching processes instead of specifically mental concepts distinct from our own decision- and action-guiding resources, it provides support for a simulationist account of mental concepts. But for now I just want to sketch the leading simulation theories of concepts.

7.3.1 Barsalou's perceptual symbols

7.3.1.1 Outline of the theory

Barsalou starts out by rejecting what he takes to be a fundamental claim of most theories of concepts, namely the notion that concepts are constituted by amodal redescriptions of modal information. These redescriptions, according to such views, occur in a "…semantic system separate from episodic memory and modality-specific systems for perception, action and emotion" (Barsalou 2003, 84). The resultant amodal symbols take the form of feature lists, semantic networks linking the features of a concept, or frames containing parameters for the essential features of the concept in question. Conceptual reasoning then acts upon these amodal representations, i.e. without having to call up memories of the original sensori-motor states.

Barsalou acknowledges that such amodal theories have distinct advantages – they elegantly account for the type-token distinction, for categorical inference, productivity and propositions, are formalizable, and can be implemented in computer hardware (Barsalou 2003 85). Nevertheless, Barsalou asserts that modal approaches constitute an attractive alternative, which has been gaining in popularity in recent years, above all due to the (putative) fact that modal approaches have greater empirical support. So what do modal approaches say?

Modal approaches dispense with re-description (or transduction), and substitute "adjacent memory systems." Barsalou illustrates the role of these systems with an example. When one sees a car, neural feature detectors are active in the visual system. Conjunctive neurons in a nearby area conjoin the active features and store them in memory. This is the "capture" procedure. These sets of conjunctive neurons also account for the trans-modal nature of concepts, namely by integrating the feature detection activity that occurred during visual perception of the car with feature detection activity that was active in other modality-specific systems, such as the auditory system. Later on, when one reasons about the car or about cars in general, the conjunctive neurons activate the neurons in the visual system and/or in other modal-specific systems that were active when the car was perceived, thereby *simulating* the sensory perception of the car. This is the "re-enactment" or "simulation" procedure.

There is an important difference between these conjunctive neurons and the associative areas of classical accounts. Conjunctive neurons do indeed associate representations, but their activity remains grounded in modality-specific systems, since they work by activating modality-specific representations. Without the occurent activity of modality-specific representations, they cannot do their job at all. If we call them associative, then we must be careful to distinguish them from purely associative areas, which would integrate or link representations on the basis of amodally coded information about their specific features, i.e. without activating the modality-specific representations in simulations of perceptual experiences. Damasio speaks here of "convergence zones", the role of which is to "…enact formulas for the reconstitution of fragment-based momentary representations of entities or events in sensory and motor cortices." (1989, p.46)

Barsalou notes that the simulations are surely not complete; they could represent a given instance of the concept, an average of instances, or a variety of other possibilities. The simulation typically involves multiple modalities. A convincing theory will of course have to be more specific about how comprehensive the simulation must be and how the parameters are selected and limited, especially since this sort of structuring of empirical input is exactly what concepts are supposed to achieve. Giving a satisfying account of how perceptual symbols differ from images, i.e. how percepts are combined and/or de-composed, is important in overcoming the limitations of imagism. Barsalou notes that re-enacting or simulating past perceptual states is not enough for a fully functional conceptual system. Such a re-enactment theory would share all the problems of imagism. (in particular compositionality and abstract concepts).

He cites two mechanisms that help to improve upon imagism: *selective attention* and *memory integration*. With the help of selective attention, the capture procedure applies not to entire scenes (e.g. a street) but to components of them (e.g. a car on the street). Thus, when focusing attention on the car, specific (modal and conjunction) neurons become active through the capture procedure. Memory integration causes this pattern of neural activation to be integrated with similar previous patterns. As a result of such cumulative integration, a perceptual symbol system for cars is formed. When, in the future, a subsequent car is perceived, this perceptual symbol system becomes active, presumably because cars have a set of features (lines, shapes, colors, sounds, contextual factors, etc.) roughly in common, and the perception of these features causes the activation of common neurons.

How to put this in terms of category identification and production? This, I think, is a strong point for Barsalou. Once a certain threshold of activity is reached in areas where there has been previously been activity in the presence of cars (capture), the *identification* of the examplar as a car (at least provisionally) occurs. Then, conjunctive neurons activate other modal neurons in the perceptual symbol system for cars (simulation) – even before the features they represent are actually perceived – and thereby giving rise to an expectation that those features (headlights, windshield, engine, etc.) will be present. This plays the role of a prediction or a category inference, or what we have called *category production*.

Empiricist theories like Barsalou's tend to have problems with abstract concepts, such as TRUTH and with concepts the content of which is specified by non-perceptual features, such as UNCLE. An uncle is not just a nice guy who teases you and gives you presents; he has to stand in a specific relation to one of your parents. Likewise, a dog dressed up as a raccoon may share the perceptual features of a raccoon, but it is still a dog. In order to recognize this, we need non-perceptual criteria and access to non-perceptual information. And it is not clear how Barsalou can incorporate such criteria or such information.

What does he say about abstraction? One: "simulators that represent the properties and relations for a category constitute abstractions about its instances (e.g. *wings* and *flies* for *birds*)" (2003, 89). What I think he is driving at is the following: simulators contain neurons that are sensitive to particular features of a concept, not to entire concepts. Hence, they represent not entire scenes or objects, but specific features. This means that they abstract from other features. This may solve the old empiricist problem with representing a triangle irrespective of the relative sizes of its interior angles. The overlap in activation common to all triangles (a.k.a. the triangle concept) includes neurons that are sensitive to closed shapes with three straight lines and three interior angles. Activating these neurons could be possible without activating neurons sensitive to certain specific features such as the relative sizes of the interior angles.

Another suspicion arises at this point. If we take Barsalou's explanation of abstraction seriously, according to what criteria are the features of an object selected that are to be activated when employing the concept of that object in thought, i.e. when thinking about the object? Take the triangle for example. Note that the features mentioned (closed shape, three staright lines, three angles) are exactly the features a definition would contain. If we allow Barsalou to say that activating just these features is sufficient, then it looks like he is just telling us how definitionism is instantiated. After all, by choosing these features he has distanced himself from imagism, since there is no easy way to make sense of the notion of an image of a triangle that has just these features and no others. In short, it seems like Barsalou's theory may, surprisingly, slide into definitionism, albeit with the difference that tokening features form the list of necessary and sufficient conditions involved modal representations for Barsalou. Maybe this is not so bad. But, then, it is especially paramount to find out what the evidence is for claiming this modal nature of conceptual thought.

Okay, but what about other abstract concepts? Some concepts, such as FURNITURE and TOOLS, appear not to require any particular perceptual features. So it is hard to see how some core perceptual features could be chosen to represent the concept. Other concepts, like THE FUTURE, TRUTH, or mathematical concepts, may not have any perceptual features at all. In other words, Barsalou's perceptually based theory needs to be supplemented in order to deal with abstract concepts. And Barsalou does in fact offer a few ideas about how to go about doing so, although they must be regarded as a mere starting point, and are not too terribly convincing as they stand. Nevertheless, I would like to present them anyway, so that I can try to build upon them later on in dealing with a particular kind of abstract concept that is of special interest to us, namely mental concepts.

The three primary mechanisms that Barsalou sketches for thinking with abstract concepts are: framing, selective attention and proprioceptive and introspective symbols. Framing refers to the simulation of event sequences that are temporally extended and incorporate background information. In support of the claim that background information and event sequences are important for abstract concepts, Barsalou refers to a study showing that people list more background and introspective features for abstract concepts than for concrete concepts (Barsalou, Solomon and Wu 1999). Selective attention, as we have already had occasion to mention, refers to the process by which salient features of a situation (or an event sequence) are highlighted. Proprioception and Introspection may require a few additional remarks.

Proprioception is in itself straightforward enough. And since it can be conceived as a kind of perception, it need not imply any substantial break from the empiricist tradition that tries to root knowledge in perception. Nevertheless, it signals an important shift by highlighting the role of the motor system and of our abilities and dispositions to act in perceiving and conceiving the world. Appreciating and working out the potential of this thought is a main aspect of the embodied cognition movement, which we will discuss in chapter 8 (sec. 8.3). I will later argue that one of the primary contributions of ST is to bring this paradigm to bear upon the folk psychology discussion. *Introspective symbols*, which are not as suspect a category as they may appear to anyone familiar with the history of psychology in the twentieth century, are identified specifically with proprioception, emotion states, and something that Barsalou calls “cognitive operations”, which, for Barsalou, include “rehaersal, elaboration, search, retrieval, comparison, and transformation” (1999, 585). In speaking of introspection of these operations, Barsalou appears only to mean monitoring of them, which presumably need not be conscious and is presumably also fallible. In other words, it appears appears to fall under what could also be called metacognition, broadly speaking. Metacognition will be the focus of discussion in chapter 8 (sec.8.2); suffice it to say for the moment that debates about introspection in the early twentieth century needs not be brought to bear upon this point. Taken together, the three mechanisms constitute a three-step strategy:

> First, identify an event sequence that frames the abstract concept. Secdon, characterize the multimodal symbols that represent not only physical events in the sequence but also the introspective and proprioceptive events. Third, identify the focal elements of the simulation that constitute the core representation of the abstract concept against the event background (Barsalou 1999, 600).

To see how these mechanisms may work, let’s take an example of Barsalou’s. He thinks that the core sense of TRUTH could be represented in something like the following manner. First, subjects construe a simulation of a physical scene either by reading about it or by hearing a verbal report – i.e. “The balloon is above the cloud”. Then, they observe a physical scene featuring a balloon and a cloud in some relation to each other. These are two events that make occur in the event sequence. The next step is to select and attend to the relevant features, namely the balloon, the cloud and the relation. Finally, they try to match the perceptual simulation to the perceived scene. After numerous successful experiences of such mapping, “…people learn to simulate the experience of successfully mapping an internal simulation into a perceived scene” (Barsalou 1999, 601). This generalized experience can be represented via introspection (in the sense of metarepresenting a cognitive operation, namely comparison). So the concept of truth is a simulator that successfully compares representations. FALSITY, on this account, is the generalized experience of failed attempts to fuse simulations to perceptions.

Obviously, this is a very schematic account, which as it stands cannot fully satisfy the need for an explanation of TRUTH – neither of the psychological basis of the concept nor of the philosophical term. For one thing, Barsalou does not address the question of whether this mapping might need to be innate, since some kind of basic hold on truth might be needed in order to get started learning anything at all. If so, then what perceptual simulations or perceived situations could serve as the input to the compare-operation? How could such a process work when assessing the truth of, say, mathematical statements? And, more generally, how do more sophisticated abstract concepts develop from this very fundamental basis? There are options for addressing these issues, but at the moment that must be regarded as work to be done.

Barsalou also addresses one other challenge to traditional Empiricist theories, namely compositionality. He addresses the issue in terms of the productive combination of simulations: "Productivity – the combination of existing concepts to form complex new ones – arises through the ability to embed simulations in one another" (2003, 89). How does that work? Barsalou: "Once an initial simulation has been constructed, one or more of its regions can be instantiated systematically with contrasting simulations to form new concepts (e.g. varying the contents of a simulated box)" (2003, 89). To take a simple example given by Barsalou, let's look at how the simulators for BALLOON, CLOUD and ABOVE can combine to form a simulation of the situation expressed by "The balloon is above the cloud." The simulator for ABOVE would represent two schematic regions in vertical alignment, the upper region perhaps having some sort of special marker (Barsalou 1999, 592). Of course, it must be emphasized that the neural instantiation of this simulator is not going to contain pictures of empty regions or of anything else. One must conceive of it as a group or network of neurons that is active when above-relations are selectively attended to in external scenes. The BALLOON and CLOUD simulators are then embedded in this schematic representation, which is to say that they are also activated and are indexed to their respective regions in the schematic representation.

Compositionality, as I mentioned in discussing Fodor's informational atomism, can also be taken too far: departures from systematicity also need to be explained. That is, one must account for the phenomenon of features that are *not* inherited by complex concepts from their constituents – e.g. CARNIVOROUS PLANTS do not have the features of fur and claws, whereas CARNIVORES do. Emergent properties are the pendant to uninherited properties – e.g. PET FISH are typically ascribed such features as being colorful living in bowls, whereas neither feature is typical of FISH or PETS.

One option is to suggest that some complex concepts are not really composed of simpler concepts after all, but are learned as additional primitive concepts that have a merely superficial verbal similarity to their constituents. In other words, we may learn the concept PET FISH by interacting with pet fish and not by fusing PET and FISH; then we would simply have a separate simulator for it that is not

7.3.1.2 Empirical Evidence

Barsalou claims that modal approaches are empirically well-supported, whereas amodal theories are not. He is careful to note that amodal theories can account for all the data, but thinks that they do so in an ad hoc way, and that they should carry the burden of proof, since there is no empirical evidence for the presence of amodal symbols in the brain, whereas there are of course modal neurons. Another argument he makes – along with all supporters of modal theories – is that an amodal symbol system would have to repeat the relations mapped in sensory systems and would thus be redundant (parsimony argument). Let me take a moment to report a bit of the empirical data he points to.

Typicality effects

- Features people list for a concept differ according to how they are asked to visualize the referent – e.g. while visualizing a lawn they do not mention roots for lawn, but while visualizing a roll-up lawn they do (2003, 86).
- Modality switching slows verification of features as typical – e.g. participants are faster at verifying ‘loud’ for BLENDER after rustling for leaves than after tart for cranberries (2003, 86).
- Perceptual similarity from one trial to the next speeds up verification of features as typical – e.g. verifying ‘mane’ for PONY is faster after verifying ‘mane’ for HORSE than after verifying ‘mane’ for LION (2003, 86)

Perceptual simulation in language comprehension

If subjects read about a nail being hammered into a wall, they are quicker to identify a horizontal nail than a vertical one, whereas the reverse is true after they read about a nail being hammered into the floor. This suggests that they simulate the perception of the nail as they are reading about it (Stanfield, R.A. & Zwaan, R.A. (2001).

Bodily states in conceptual processing

There is tons of work showing the relationship between the body (postures, grips, eye direction, etc.) and conceptual processing. Just a couple of examples: people tend to look up more often than down when hearing a description a rooftop (Spivey et al., 2001); they are also quicker to verify that "Open the drawer" is a sensible sentence when their response involves a pulling motion as opposed to a pushing motion (Glenberg, A.M. and Kaschak, M.P. 2003). Although Barsaou does not state this, he is obviously appealing to states of the motor system here that are influenced by conceptual thought – not perceptual states (except perhaps proprioception). I do not think he would have anything against including motor representations within simulators, but as far as I can tell he does not himself do so. I, on the other hand, think it is a legitimate and important addition that should add power to the conceptual system, and also helps us to see how MNs can be integrated in action understanding using concepts.

This data does uncontroversially demonstrate that conceptual thought can at least be affected by modal and motor representations and vice versa. But this does not necessarily imply that conceptual thought takes place exclusively or even primarily in modal systems. So it is open to a skeptic to stick to the position that conceptual thought is amodal but is not wholly independent of modal systems.

7.3.2 Prinz's Proxytpe Theory of Concepts

Prinz (2002) builds upon Barsalou's theory, attempting to work out its philosophical significance and also to improve on some of its limitations. In particular, he wants to explain publicity, intentionality and compositionality within Barsalou's framework (2002, 152). Abstract concepts and are an additional challenge that he addresses. As we will see, they remain (unsurprisingly) the weak point for perception-based approaches.

Prinz formulates the basic principle of his concept empiricism as follows: "All (human) concepts are copies or combinations of copies of perceptual experiences" (Prinz 2002, 108). This thesis is intended to be limited to the nature of mental representations or vehicles of thought, thus remaining neutral about epistemology and semantics. Prinz's basic move in positioning Barsalou's theory within the landscape of philosophical conceptions is to start out (perhaps surprisingly) by borrowing Fodor's account of intentional content. He endorses the information-theoretic account of intentional content[82], but argues that concepts are not to be thought of as indicators but as detectors in the sense outlined above in section 7.2.5 (Prinz 2002, p. 124-127).

Note that this twist immediately skirts the problem with acquisition that I discussed at the end of section 7.2.5. Fodor has no way of accounting for how atomic symbols come to be linked with the right intentional content in such a way that everyone shares the same concepts. Prinz has no problem with this, since his concepts are derived from perceptions. Perceptions are stored in long-term memory networks, as described by Barsalou, and re-activated when the concept is invoked. Their link to the object outside is therefore not arbitrary but determined by the properties of the object. Prinz can therefore combine an elegant explanation of acquisition with the information-theoretic account of intentionality endorsed by Fodor.

Cognitive content also comes rather easily on a perception-based approach such as Barsalou and Prinz espouse. One need only note that perceptual systems detect appearances, i.e. perceptible features of things, which can differ over time, in various settings, etc. So, the morning star may appear a bit different than the evening star. Hence, the perceptions that are reliably caused by them may differ a bit. So there is room for cognitive content to depart from intentional content. I will have to say a bit more about the details of Prinz's theory before I can discuss his explanation of cognitive content.

Proxytypes

Prinz's proxytype theory is motivated in part by a particular problem he raises for the view that concepts are the constituent of thoughts. Consider the distinction between occurent thoughts and standing knowledge. I may have the knowledge that Paris is not the capital of Romania without currently tokening that thought. Thoughts are stored in working memory for their duration whereas standing knowledge is stored in long-term memory. If concepts are the constituents of thoughts, what happens to a concept when one ceases to token it within a thought? Does it just vanish? Do I cease to have the concept of a dog when I am not thinking about petting one or running away from one? This seems odd. Moreover, concepts do seem to need to involve a whole array of background knowledge about a category, such as typical features, relevant causal or theoretical knowledge and so forth. So it is reasonable to

[82] In fact he seeks to improve on it a bit, arguing that the link between a concept and its intentional content is fixed not only by a causal chain or a nomological relation but also in part by the learning history of the individual concept-user. This enables him to exclude some cases that are problematic for causal accounts. But for the purposes of the discussion here, we can ignore this detail.

incorporate a link to background knowledge stored in long-term memory in one's theory of concepts.

Prinz introduces the term proxytype in order to explain how this can be done. Proxytypes are "mental representations of categories that are or *can be* activated in working memory" (2002, 149). These mental representations, on this empiricist view, are perceptually derived. To put this into context, proxytypes are equivalent to simulators in Barsalou's theory. Not surprisingly, Prinz also avails himself of the concept of simulation, saying "if concepts are proxytypes, thinking is a simulation process" (150).

The representations making up a proxytype can be mono-modal, such as a visual model of a dog, or they can be mulit-modal, incorporating auditory representations, such as a dog's bark. Prinz is clear that they can also include words, since words are of course also visually or auditorily perceived. Moreover, they can contain theoretical knowledge. This helps to deal with some of the data that has been used to support theory theory, such as Keil's painted raccoon example.

The multitude and diversity of representations that are available in long-term memory and can be actualized in thinking about, say, a dog, means that the dog concept can be tokened in a lot of different ways, depending on what perceptual information I am presently confronted with or on the context (i.e. when hearing a story about the arctic tundra I may token a sled dog, whereas when I am being told about a guard dog I may token a different breed). This raises a potential problem for the *publicity* of concepts – like the problem that imagism has with publicity, but even worse since it undermines the stability of a concept even within an individual.

A first point of qualification to make is that concept sharing may be overrated. That is, there is empirical data revealing that people's typicality judgments actually vary considerably, even the same person's judgments from one occasion to the next (Barsalou 1989, 1993). But still, we need an explanation of the substantial publicity of concepts that must obtain. In order to give such an explanation, Prinz suggests that there are such things as default proxytypes, which are simply the proxytypes that we token in the absence of a context. The features that go into a default proxytype are the features that one must frequently represents a category as having. This frequency is a function of various factors, including perceptual salience, cue validity (subjective probability that a thing belongs to a category given that it has a feature) and conformity to relevant theories.

Default proxytypes also explain *typicality effects* quite easily: typical features are simply the features included in the default proxytype (Prinz 2002, 162). The same goes for *basic-level categorization*. Prinz point out that basic-level categorization involves that kind of schematic shape representations (one shape suffices for beagles, huskies and rottweilers) that our perceptual system most likely process best: we discern gross shapes more efficiently than fine details (Prinz 2002, 163) One could expand on this, pointing out that other perceptual representations would be fairly constant across various specific breeds of dogs, such as barking and stinkiness.

Beyond this, I think one could enhance this explanation by adding motor representations into proxytypes. As we will see in chapter 8, on mirror neurons, it has become increasingly clear in recent decades that the motor system is involved in perception (i.e. not just in action). For the moment, suffice it to say that perception of objects (e.g. dogs) activates motor representations of ways of interacting with that object. Presumably the ways that I interact with dogs is pretty similar regardless of specific breeds: petting, hiding my face from errant tongues, running away, etc. Hence, motor representations may contribute to the bias toward basic-level categorization.

With respect to compositionality, there are a couple of things to say. Firstly, Prinz points out that predicting too much compositionality may also be a problem for a theory, and asserts that Fodor's informational atomism may be in danger of doing so. Prinz gives an example involving the combination of three concepts: OSCAR, APPLE, and EAT. Fodor's theory predicts that "Oscar eats the apple" should be no easier to understand (i.e. no quicker to process) than "The apple eats Oscar", since the two nouns have the same syntactic features. But it seems likely that there is in fact asymmetry here, which Prinz would explain by appealing to the visible features of the objects in question, such as Oscar's having a mouth, which is useful for eating, whereas apples have no mouth (Prinz 2002, 300). I think Prinz is right about this, and some empirical work testing such predictions would be highly welcome.

Prinz thinks that a lot of compound concepts probably are stored in long-term memory networks just like simpler concepts, having been learned through experience with the objects in question rather than through combining the simpler concepts. PET FISH, for example, are encountered often enough that one's concept of them is likely relatively independent of the concepts PET and FISH. The suggestion is that compositional processes only come into play when compound concepts are not retrievable from memory. Prinz proposes a couple of procedures that could play a role: "aligntegration" (i.e. alignment + integration) and "feature pooling".

To see how aligntegration works, let's consider an example discussed by Prinz[83]. To interpret the compound SQUIRREL SNAKE, one seeks a parameter of the SNAKE concept with which SQUIRREL could be aligned, and hits upon the DIET parameter. The default setting might be MICE, for example. Since mice and squirrels are pretty similar, SQUIRREL is aligned with the DIET parameter. The integration step has to do with the adjustment of feature weights. SNAKE involves various parameters for various features, some of which are more important than others and are therefore weighted higher. In our example, the MICE setting of the DIET parameter may not be all that crucial, since snakes may eat birds or other lizards. But once SQUIRREL is aligned, the importance of the DIET parameter and thus the feature weight must be raised, since a squirrel snake is only a squirrel snake if it eats squirrels.

What about abstract concepts and other theoretically derived concepts? This is truly the major challenge to Empiricist approaches such as simulationism. I tried to show above that Barsalou's theory gives us a good way of dealing with at least some concepts that are abstracted from perceptions, such as the concept of a triangle (independently of the relative sizes of its interior angles), and at least some ideas about how to deal with concepts that resist characterization in terms of any particular perceptual features at all. Prinz also contributes similar speculations about how to construct abstract concepts from perceptions. Let me mention four strategies that Prinz presents for accounting for abstract concepts.

Firstly, Prinz proposes a mechanism he calls *tracking*. (Prinz 2002, 170) Although we may not always have perceptual access to the features that are decisive for falling under a concept, it may often be the case that we have perceptual access to features that are correlated with – i.e. track – the decisive features. For example, regular succession famously tracks causal relations pretty well, and may form the basis for a perceptual representation of causality. Admittedly, this does not exhaust the concept – not even the intuitive, everyday concept, which involves an aspect of necessity and perhaps also transference of force. Prinz therefore suggests that our everyday bodily experience of acting upon objects, including our own bodies, and being acted upon, may also be incorporated in the proxytype for CAUSALITY (177). Now, obviously, this is pretty speculative. He does review some developmental data, but my intention here is only to display how tracking can be supplemented with other representational resources to form an abstract concept. Note that the addition of bodily experience is in line with Barsalou's addition of proprioception.

Secondly, *verbal* skills can also contribute in numerous ways to the formation of abstract concepts. It may seem illegitimate for an empiricist to fall back on linguistic

[83] Prinz borrows the example from Wisniewski (1997).

representations. But I do not think that is the case. After all, language is not – at least not primarily – represented amodally, but auditorily, visually, and in motor representations of one's own speech. The most obvious way is simply by remembering verbal associations. One may be able to infer (i.e. recall the association), for example, that something is negatively charged if it is an electron, even though one has never perceived an electron and does not have a very definite idea about what it means to be negatively charged.

Another point, not raised by Prinz, could be made about research in cognitive linguistics about the relationship between *grammar* and conceptual structure. In a commentary on Barsalou, Langacker points emphasizes that cognitive linguists regard grammatical constructs as meaningful (i.e. having semantic content) insofar as they impose construals upon conceptual content (1999, 625). For example, one and the same object (e.g. dad, man, guy with the Grateful Dead t-shirt, etc.) can be referred to with various nouns, depending on the aspect one wishes to draw attention to. Different verbs can also be used to describe the same event, depending one's interests and on one's evaluation/interpretation (e.g. running, fleeing, hurrying, etc.). By the way, Barsalou also considers linguistic communication, in particular communication about past, future or otherwise currently absent situations to be a kind of simulation (1999, 594).

A related contribution of cognitive linguistics to our understanding of concepts, which is mentioned by Prinz and also by Barsalou, lies in the investigation of the role of *metaphors* in conceptual thought. George Lakoff (Lakoff & Johnson 1980) has contributed especially to our understanding of the ways in which metaphor structures conceptual thought. The basic idea is that we use representations of perceptually salient entities, processes and events in order to think about abstract entities, processes and events. Hence, our thinking about ANGER metaphorically projects perceptions in a container, i.e. it heats up to a boiling point, increases in pressure, explodes, seeps out, etc. The container-metaphor is typical for mental states, according to Lakoff and Johnson (1980), and structures our thinking about communication (transference from one container to another) or thoughts and knowledge (contents of a container), influencing, for example, the inferences we are inclined to draw. This is all quite plausible, and my very brief sketch does not do justice do Lakoff and Johnson's detailed proposal, nor to the interesting empirical work that has been conducted since then, but I will return to this issue in the next chapter, namely in section 8.3 (Zhong, C. & Liljenquist, K. 2006; Zhong, C. & Leonardelli, G., 2008). But, as Barsalou points out (1999), metaphoric projection needs to be complemented by procedures for identifying which features are to be projected (i.e. which features are similar or the same between the two relata) and which are not.

A fourth option is the appeal to *operations*. Some abstracta, such as NEGATION, may be regarded as operations rather than as concepts. What Prinz seems to have in mind is something like Barsalou's rough-and-ready explanation of TRUTH and FALSITY, discussed above. But Barsalou does not make the move to stop calling these concepts concepts; rather, he thinks that some concepts include dynamic operations. I think this is probably the wiser position, since otherwise we would have to revoke the appellation from a whole lot of concepts that include dynamic manipulations of representations.

One other important point that Prinz contributes to the discussion of abstract concepts is that informational atomism is in exactly the same boat as Empiricist approaches (Prinz 2002, 167). The reason is simple. Take an abstract concept such as DEMOCRACY. You cannot taste, smell, hear, touch or see a democracy, so simulationism has a problem. It may seem that informational atomism has no problem, since its symbols stand in an arbitrary relation to their intentional content. But in order for these symbols to have intentional content, information atomism proposes that they stand in a nomological causal relation with things outside. But if the symbol for democracies gets its meaning from a causal chain starting with an encounter with a democracy, then it must be possible to perceptually detect democracies, or their manifestations at any rate.

This may not come as a surprise when we reflect that Barsalou and Prinz both explicitly adopt the causal account of intentional content that forms the basis of informational atomism. It shows, at any rate, that Empiricism's problems with abstract concepts do not arise specifically from Empiricism but from causal theories of intentional content. This is relevant for us insofar as understanding actions can require understanding abstract intentions, i.e., goals which have a certain autonomy vis-à-vis the specific movements used to realize them. are (at least prima facie) abstract. We will there First another simulationist theory, then in th next chapter some ideas that help with ascription, the other problematic component of action understanding.

7.3.3 Embodied concepts: Gallese & Lakoff

Since one of my aims in discussing simulationist theories of concepts is to contribute toward a simulationist interpretation of MNs, it makes sense to conclude this chapter on simulationist theories of concepts by reporting on Gallese's views on this matter. Interestingly, Gallese constitutes the one exception to the surprising rule that advocates of ST do not mention simulationist theories of concepts. Gallese has in fact engaged acttively in the discussion about simulationist theories of concepts, and has written papers together with Thomas Metzinger and George Lakoff. Gallese regards concepts as being embodied, which appears to mean that they are constituted within the sensory-motor system (see Gallese 2003; Gallese 2005; Gallese & Lakoff 2005, Gallese and Metzinger 2003). He also says that they need not be conceived as "linguistic" entities (Gallese 2005). What could this mean?

Gallese apparently concurs that, in order to get a representational structure that is sophisticated enough to count as being conceptual, the motor system is not enough, since it underdetermines one's intentional state. Gallese must therefore add something while stopping short of abstract symbols. What he seems to have in mind is a view of concepts as being constituted by a combination of motor representation *plus* various kinds of perceptual information, but without additional symbolic representations. Gallese also says elsewhere that "conceptual knowledge is structured and its content determined by the activation of neural clusters within the sensory-motor system" (Gallese 2005).

In order to improve upon Gallese's account, I would like to start out by simplifying matters a bit by putting mental concepts aside for a moment and talking first about object concepts. Then I will move on to one's own goal concepts, which presumably contain object concepts. You may have reservations about my distinguishing them from mental concepts, since they prima facie still require an understanding of the distinction between the world and mental representation of the world. True enough, but it is a different, and simpler, kind of representation that is at stake in first-person goal representation as opposed to ascriptions of goals. A representation of one's own goal has a world-to-mind direction of fit. Ascription of a goal state, on the other hand, is itself not a goal state but a belief state; it has a mind-to-world direction of fit. In fact, insofar as it seeks to accurately represent the other person's goal state, it is a belief state that contains a goal state within its content. As such, it is structurally more complex. Hence the sequence: object concepts, action concepts, mental concepts

7.3.3.1 Embodied Object Concepts

It is important to be clear about what is meant by the claim that the content of conceptual knowledge is determined by neural activity in the sensory-motor system. One could read this as a fairly innocent affirmation that different conceptual contents correspond to different patterns of neural activation. But the content could, in this case, nevertheless be a set of features, a set of objects that fit the bill, a Fregean sense, or just about anything else. But this misses the point suggested by the specification that the neural activation is in the sensory-motor system. What this notion actually suggests is that the content of a concept crucially involves actions that can be performed with its exemplars. Gallese speaks here of the “relational aspects of categorization processes”[84].

He illustrates this notion with a few examples, one of which is a series of studies that show that observing or silently naming tools, or even imagining that one is using them, leads to activation in the ventral pre-motor cortex. Gallese finds this cool because he thinks it reveals that relational specifications of tools “comprise a substantial part of their representation” (Gallese 2003, 1235). Gallese also refers, of course, to MNs and canonical neurons. With respect to the latter, he discusses the finding that a “considerable percentage” of them are active when performing a particular grip or when observing objects that, although different in appearance, afford that same grip. Gallese concludes that these canonical neurons contribute to a multi-modal representation of an “organism-object relation” (1236).

The content of these representations, Gallese argues, must include a specification of the actions that are preformed upon the objects in question. It is, as he puts it, “the result of the ongoing modeling process of an organism as currently integrated with the object to be represented, by intending it” (1236). The upshot of this line of thought is that the content of a perceptual object representation, and therefore also the content of an object concept, is not the object per se, but a bunch of perceptual features of the object and a bunch of actions that involve the object. Gallese’s use of the term modeling may appear a bit strange here; it may seem more natural to say that he is doing away with the internal model of the object that is extracted from perception, and replacing it with probabilistic links between sensory and motor processes. The translation between these systems does not need, for him, to be mediated by a model at all. In fact, I am not sure that it is legitimate to speak here of representations. If one does so, then the representation is distributed throughout the entire body. “The producer and the repository of representational content is not the brain *per se*, but the entire organism, by means of its interaction with the world of which it is a part” (1237). This is an implicit appeal to the idea of virtual models espoused by Noe (Noe 2001), O’Regan, (Noe and O’Regan 2001), and other advocates of embodied cognition, which will be discussed briefly in chapter 8.

But getting back to the point: is Gallese’s interpretation of his data persuasive in this regard? One might object that the activation of these areas does not show that they play a role

[84] Gallese 2003, p. 1235.

in categorizing the objects; the activation may occur subsequently, with possible actions being considered as a result of the categorization. Alternatively, one may object that it is not surprising that relational features play a role in specifying the content of some concepts, such as tool concepts, since what makes tools into tools is not primarily their appearance but their usefulness for particular actions. It may be that relational aspects are part of the content of tool concepts but that they constitute an exception rather than the rule in this respect.

Gallese also gives another reason why it is reasonable to think that the activity in motor areas during these tasks is part of the categorization process – namely, because it makes more evolutionary sense. Evolution, he asserts, is parsimonious. Higher-order cognitive processes are likely to be built onto and to make use of sensory-motor circuits, which are, after all, responsible for controlling the activity of the organism. And higher-order thought, if it is to make any difference for survival, must modify the organism's behavior. That is probably so, but it is unfortunately nothing more than a just-so story.

In any case, this theory of concepts suggests how, for Gallese, MNs can partially constitute conceptual understanding, since conceptual understanding involves or even is simulative. Gallese goes so far as to say that conceptual understanding is imagination, and that imagination is a kind of simulation. Gallese and Lakoff write: Consider a simple sentence, 'Harry picked up the glass.' If you can't imagine picking up a glass or seeing someone picking up a glass, then you can't understand that sentence" (456). If MNs constitute simulations or parts of simulations, then they can on this view at least partially constitute conceptual understanding. Let's see how this works for action concepts.

7.3.3.2 Embodied Action Concepts

Introducing actions into the picture is an extension of the same strategy. It seems that the combination of sensory and motor activation could *function* as a representation of an intention to act without having to give rise to *any additional representation* beyond the representation of the object's properties and the planned movements. These representations (or models, or virtual models, or whatever) simply constitute a different representation with a different content when the overall situation in the brain is different. Gallese and Metzinger (2003) write, for example: "to speak of a neural level of representation that *has* to be qualified as 'goal-related', however, doesn't necessarily imply that this level *per se* is also able to specify the action goal in a way that is *detached* from the means to achieve it." What Gallese and Metzinger want to do, then, is to localize action concepts in the sensory-motor system to as great an extent as possible.

The basic idea that Gallese and Lakoff (2005) present is that “the job done by what have been called ‘concepts’ can be accomplished by schemas characterized by parameters and their values.” (467) A schema, on their view, consists of a “network of functional clusters.” Each cluster characterizes a particular parameter. One cluster for each parameter value. Parameters, for Lakoff and Gallese, are higher-level features of neural organization, constituted by firing frequency and level of activation of particular groups of neurons. So, depending upon the way in which a particular bunch of neurons are behaving, you get different settings of the parameters associated with those neurons. That parameter could be action type, intensity /level of force), direction, object location, agent (self or other). This last is especially important, since the self/other distinction would seem to play a central role in ascription. If MNs are symmetrically active in action and perception of action, we still need to account for how the activity is understood to be my own action in the one case and someone else’s in the other. We will come back to it below, when we deal with ascription.

The important point is that for Gallese, Lakoff and Metzinger, an action is constituted by a combination of motor and perceptual representations, and not by any additional symbolic, amodal representation. Representations of actions are not just MN activity, nor are they additional representations alongside the motor representation and whatever perceptual representations may be relevant, but a unique combination of them. One feature of this proposal is that the contents of the various representations that constitute the representation of an action are states of affairs in the world, not in the agent’s own mind. But we still have to introduce the element of ascription that appears to be necessary for understanding others’ actions, that is, to differentiate between understanding what one is doing (or imagining oneself doing) and understanding what someone else is doing?

7.3.3.3 Embodied Mental Concepts (Ascription)

One choice presented Lakoff and Gallese is agent parameters. The idea is at bottom a simple one. Action concepts involve various parameters, such as level of force, direction of motion, and object location, etc. And one of these parameters just happens to be an agent parameter. Gallese and Lakoff (465) mention the supplementary motor area (SMA) and the primary motor cortex as areas in which there is activation in the case of one’s own performance of an action but not in the case of action observation.

It may appear at first glance like a cop-out to simply add on an agent parameter to account for ascription. It seems like they are explaining something by means of itself. A skeptic would simply respond that mental concepts are postulated precisely in order to determine how to set the agent parameter, if we want to speak of agent parameters at all. Moreover, the proposal is not very detailed. But I would maintain that there is some meat to the proposal and that it substantially differs from what a theory theorist or someone favoring an amodal concept theory would be inclined to think.

Recall Gallese's remarks mentioned in chapter 6 (sec.6.6.2) about ascription being a default function of MNs. What Gallese and Lakoff say is an expansion of this idea and takes on its true significance if combined with this point. What they say is that activation of the SMA and of the primary motor cortex are *additionally* present when I perform an action myself. Other areas, like the MNs, remain the same in both cases. Buccino et al. (2001) Jackson and Decety (2004).

So it does seem possible to distinguish between representing my own action and representing someone else's action with the help of an embodied theory of concepts in which MNs feature prominently. But, crucially, MNs are not responsible for this distinction. The question then arises whether they can be said to be representing the action as an action although they do not appear to be representing the agent and are therefore not doing the ascribing? Obviously, Gallese wants to affirm this. Pacherie and Dokic (2006), however, argue that it does not help to say that they represent the action as being performed by an agent but leave the agent parameter undetermined, because the criteria for applying the concept of representation include, among other things, the possibility of contrast:

> A state represents an aspect of the world only if it is produced by a system that is also capable of producing representations of contrasting aspects. Something is always represented, at least implicitly, as opposed to something else (Pacherie and Dokic, 2006, 103).

If MN activity does not differ between my own movements and observed movements of someone else, then it does not contain this possibility of contrast, and thus does not represent an agent at all. And since all of the participants in the debate seem to agree that action understanding involves representing not only a movement but also an agent and an object, MNs' failure to represent any agent at all disqualifies them as representations of actions.

But wait just a minute. There is still at least theoretically the possibility that MNs distinguish between movement of an agent and movement without an agent. Of course, it is no straightforward matter how to decide this, since observed movement has to be movement of something, and that something could be regarded as an agent. Hence, whenever there is movement there is an agent. One solution to this difficulty would be to say that inanimate shapes, points of light, lines and the like are not biological agents even if they are made to move as though they were. If MNs do not respond to these movements, then they could be said to distinguish between movements of agents and movements in the absence of agents – and therefore to represent agency even if they cannot represent specific agents. Of course, this would not necessarily mean that they are responsible for inferring that there is agency in the one case and not in the other. That could be accomplished elsewhere. The only demand is that they reflect this distinction.

7.4. Summing Up

We have seen that simulationist theories of concepts can help to elucidate the role of MNs in action understanding. They enable us to introduce concepts into a robust interptretation of MN activity. One effect of this is that they show how a version of ST is possible that falls between Goldman and Gordon on the question of mental concepts. We can agree with Goldman that mental concepts are invovled in simulative understanding, but we can stop short of saying that they have to be added on to the simulation process, which itself is distinct from the ascription of an intention (or beliefs and desires) to a targer person. Instead, the simulation (in some cases constituted by MN activity) *partially constitutes* conceptual understanding of a target person's action.

Note that the kind of simulation that is involved in conceptual reasoning in general for Lakoff and Gallese, as for Prinz and Barsalou, is not simulation of a target person but simulation of perceptual and motor experiences involving exemplars of the concepts at issue. In this respect, this notion of simulation is broader than the notion of simulation that is directly relevant to ST, i.e. it refers not to social cognition so much as to cognition generally.

We also saw that if we want to bring a simulationist theory of concepts to bear upon action understanding/social cognition, we have to specify what kind of picture of *mental concepts* it gives. In addressing this question, we discussed the idea of agent parameters. In order to flesh out this idea, I would like to present (chap.8) another area of research in which it fugures prominently – namely, research on action representation. Aside from clarifying the idea of agent parameters, the work I will discuss can help to clarify futher in what way the concept of simulation needs to be expanded in order to incorporate simulation *of one's own actions* as a basis for simulating others' actions.

Chapter 8:
Expanding the Concept of Simulation II:
Simulationist Theories of Action Representation

8.0 Introduction

In this chapter, I will be discussing recent work that prominently employs the term "simulation" in a way that is importantly similar to the use introduces in chapter 7, namely to refer to simulation of one's own actions. To be precise, it is not just one specific are that I will be presenting but, rather, a family of areas of research that can be grouped together on the basis of a common conception of simulation. In the first section (8.1), I will discuss work on action planning, action imagination and ascription of agency (e.g. Jeannerod, Pacherie). As we shall see, there is a sense in which we can be said to model, or simulate our actions when we plan them. There models or simulations are of use to us as points of comparison as we carry out the action in question, because we can compare proprioceptive feedback and perceptions gained during the action with the expectations of proprioceptive feedback and perceptions that the model or simulation suggests. If we also simulate others' actions in perceiving/interpreting them, this creates the problem of distinguishing between our own and othters' actions. For this purpose, Jeannerod and others invoke the concept of the who-detector, which is essentially the same as the agent parameter I have already discussed.

After discussing this work, I will look at work on metacognition (8.2), which can help to understand the kind of understanding of our own actions and of simulations of our own actions that is suggested by the work discussed in section 8.1. Since this understanding is presumably unconscious and is arguably not metarepresentational, it suggests a deflationary construal of the introspective access to mental states that Goldman introduces, which I hope will help to work out a middle path between the versions of ST espoused by Goldman and Gordon. I conclude the chapter with some reflections upon the expansion of the concept of simulation in chapter 7 and 8. I suggest that this expansion serves to place ST roughly within the context of the embodied cognition movement. The benefit of this contextualization is to show how the concept of simulation has been empirically fruitful over the past 2-3 decades, and also to reveal that it has the potential to tie together research in different areas of neuroscience, cognitive science, psychology and philosophy, and thereby to increase theoretical coherence in these areas.

8.1 Jeannerod's Simulation Theory of Action Representation

8.1.1 Covert actions

One obvious way to extend the concept of simulation beyond cases in which the mental states or processes of one person matches the mental states or processes of another person is to apply it to first-person imagined or recalled actions. There are indeed approaches in the neurosciences that make just this move. In particular, Marc Jeannerod (2000) advocates a simulation theory aimed at what he calls covert actions, or covert stages of action, by which he means planning and intending actions that will be executed – i.e. will become overt – but also assessing whether particular actions are feasible, imagining actions, recognizing tools, learning by observation, as well as understanding others' actions (Jeannerod 2000, 103). The basic idea is that there is an overlap in activity in neural networks that include the motor system between cases of covert and overt action. This overlap is also referred to as a "shared representation".

It is plain that this idea, much like simulation theories of concepts, constitutes an extension of the basic idea of ST in folk psychology, namely that the same resources or processes are used in performing, planning or deciding upon actions, on the one hand, and thinking about them on the other. The difference, of course, is that the *thinking about them* here applies not only to thinking about others' actions but also to thinking about actions that one has performed or may perform in the future. Jeannerod's motivation and his background, however, do not lie in the debate on folk psychology, however, but in the neurosciences. Like MN research, this sort of simulation theory can therefore be said to constitute independent empirical evidence for ST. Working out which version of ST profits most, again, is another issue. In order to clarify the idea, let me present some of the empirical data upon which it is based.

First of all, some neurophysiological data. FMRI studies conducted by various researchers with various setups have consistently revealed that there is an overlap in the pixels that are activated in the primary motor cortex during performance of particular movements, on the one hand, and when one imagines oneself carrying out the movement on the other (Rot et al., 1996; Porro et al., 1996, Lotze et al. 1999). Further downstream, in the muscles involved in performing an action, TMS studies have found that during imagination of actions, there is an increase in the motor evoked potentials (MEPs) specifically in those muscles that would be involved if the action were really performed (Fadiga et al., 1995 and

1999). This motor outflow in imagined action is also associated with various other kinds of activation that are characteristic of action performance, such as increased heart rate (Decety et al., 1993) and respiration, both of which respond proportionally to the imagined effort (Wuyam et al., 1995). Jeannerod (2000) also reviews evidence of an overlap in activation other areas of the motor system, such as basal ganglia, cerebellum and premotor cortex, as well as in associated cortical areas, but I will not discuss any physiological evidence here.

Apart from the *physiological evidence* for shared representations, there have been interesting experiments intended to provide *behavioral evidence* that there is an overlap in procedures and/or resources between action execution and imagination. Specifically, the hypothesis they are aiming to support is that "that mental simulation of action is assigned to the same motor representation as preparation and execution." (Decety, 2002) These studies usually work by showing that it takes the same amount of time to imagine doing something as it takes to actually do it.

One classic study that is often referred to shows that it takes people the same amount of time to walk to a target as it takes them to imagine themselves walking to the target (Decety et al. 1989). A subsequent, more refined study in the same vein showed that the same speed-accuracy tradeoff holds for simulated actions as for real actions (Decety and Jeannerod, 1996). Subjects in this study had either to walk along a path and pass through a bunch of gates or imagine doing so. Some of the gates were narrower than others and therefore demanded greater precision. Unsurprisingly, the subjects walking along the course took longer to pass through the narrower gates. As it happened, the subjects who were only imagining themselves negotiating these harrowing gates showed the same pattern.

There is one thing that holds me back from making too much of these studies. If you ask subjects to imagine performing a task, it is natural that they will go through it in their imagination as accurately as possible, taking care even to imagine it at the right speed. In other words, you are implicitly asking them to imagine performing the action at the same speed as they would actually perform it. For this reason, I find a slightly indirect kind of experiment more persuasive, in which the subjects are not asked to imagine performing an action, but are led to do so in order to complete some other task. Parsons et al. (1987), for example, showed that when a photograph of a hand was presented and subjects were asked to determine whether it was a right or a left hand, the time it took them to answer depended on the orientation of their own hand. The longer it would have taken them to move their hand into the position of the hand in the photograph, the longer it took them to answer. In a similar, more recent, study, subject are asked to assess the feasibility of a certain action, e.g. to assess whether they can grasp an object placed at different orientation. Again, the time it takes them to respond depends on how far their hand currently is from the object (Parsons 1994, Frak et al. 2001). I just want to note that there have been neuro-imaging studies revealing an overlap in activation in the premotor cortex between performance of actions, on the one hand, and these cases of assessing whether the action is feasible on the other (Parsons et al., 1995).

But, aside from the overlap, we would also expect to find some differences in the activation of brain areas depending on whether an action is performed, observed, recalled, imagined, etc. Otherwise there would be no difference between action and action understanding, which can't be right. Indeed Jeannerod presents evidence of such differences. I will put off the details until I discuss who detectors and agent parameters below. At the moment I just want to point out that the general strategy accords with the general result that has been emerging in several related discussions (MNs, ST), and which I have been supporting, namely that the basic idea of ST needs to be supplemented with provisions for integrating contextual information and performing ascription, oat least sometimes with mental concepts.

8.1.2 Ascription of simulated actions

According to Pacherie and Jeannerod (2004) MNs this empirical evidence of an overlap in neural activity in performance, imagination and observation of action has far-reaching theoretical consequences for the philosophy of mind. On their view, it reveals that differentiating between our own actions and those of other people is more problematic than has been traditionally recognized in philosophy. The negative side of this is that in identifying our own actions, i.e. identifying ourselves as agents of our own actions, we are not immune to error. In other words, the immunity to error through misidentification (IEM) does not obtain with respect to the identification of the agent of our own actions. The positive flipside is that the problem of other minds is simplified insofar as our access to others' intentions turns out to be more direct and more like our access to our own intentions than has traditionally been realized. Let me explain how this works. It helps to contrast the view Jeannerod and Pacherie present with what they present as a traditional, Cartesian view.

For Descartes (at least as they present him), bodily movements are just bodily movements. They do not have any intrinsic mental properties like intentions. Some of them constitute actions by means of the fact that they are caused by intentions. Intentions, then, are causal antecedents that precede bodily movement. We have to infer on the basis of others' movements that they are caused by intentions, and, thus, that others have minds. In the case of our own actions, however, we cannot err. If I believe that I am the author of an action, say raising my arm, then I cannot be mistaken in the ascription, since the ascription concerns a mental event, namely an intention, and my mind is transparent to me.

I am not sure that this construal of Descartes is quite right. For Descartes, the mind is indeed transparent; we can be sure what our thoughts are and that they are ours, etc. But he is not directly concerned with actions, which crucially involve bodily movements. If I have the intention to raise my arm and also see my arm ascending, it is possible that an evil demon is actually causally responsible for the movement and my intention is causally inert. Hence, I think Descartes ahs to say that I am mistaken in ascribing the agency of the action to myself. What I cannot be mistaken about – in Descartes' view – is that I have the intention. I may be wrong in thinking that it is really my arm, or I may be wrong in thinking that it is my intentions that is causing it to ascend, but I cannot be wrong in thinking that I have the intention to raise it.

On this roughly Cartesian view, the case of second- and third-person ascription is crucially different. We are not directly aware of their mental states (such as intentions to act) as we are of our own but, rather, we infer them on the basis of our perceptions of their bodily movements. Second- and third-person ascriptions of intentions to act, on this view, proceed in two steps, each of which is prone to error. We could of course be in error about our perceptions of their bodily movements, or we could err in drawing the particular inference that we draw.

Pacherie and Jeannerod, on the other hand, have a very different view of this process. According to them, we can in fact directly perceive others' intentions. The basic idea is that our perception of certain stimuli – namely, of other people performing actions – proceeds by means of our entering into neural states that we enter into when we ourselves are performing those movements with intentions. These neural states, which are common to action performance and action observation, are called "shared representations". Because of this direct link from observation of bodily movement to activation of a representation of an intention to act, we do not need an inference in order to ascribe an intention to them. Rather, the intention is a property of the physical motion that we perceive.

The plus-side (making others' minds more directly accessible) is obvious, but a problem also arises, as Pacherie and Jeannerod note. If I become aware of others' intentions via the same representation that constitutes my own intention to act, then the activation of this representation can not serve as a clear indicator of my own intention to act. The fact that it is activated does not unambiguously reveal whether I am performing an action or observing one. Pacherie and Jeannerod assert therefore that in both cases (e.g. opening a door or watching someone else open the door), I am primarily aware of an unattributed or "naked" intention to open the door, and attribute the intention on the basis of additional information (i.e. who detectors,or agent parameters)

And this is no merely theoretical possibility, as it turns out. Pacherie and Jeannerod cite a rich body of experiments revealing instances when one can represent an action but be mistaken about whether one is the author of that action. In one experiment (van den Bos and Jeannerod, 2002), for example, participants performed movements with their hands, which they could not see directly, but which were videotaped. Participants were shown a videotape of their own hand movement (while performing it), and also an alien hand performing either the same or a similar movement. The participants indeed had difficulties making the correct identification in cases where the hands where the movements were similar but not exactly the same, and these difficulties tended to be compounded when the images of the hands were rotated either 90° or 180°. These findings support the idea that distinguishing between one's own actions and those of others is not a trivial affair, and that misattribution is possible. It is also worth noting that participants tended to over-self-ascribe rather than under-self-ascribe. With respect to the general claims advanced by Jeannerod and Pacherie, the point is that the participants represent a "naked intention" and then have to attribute this intention either to themselves or to someone else.

As for the mechanisms that are involved in making an attribution, the results of the experiment suggest that visual information and proprioception both play a role, and that visual information can override proprioception. This is in fact a well-known and robust phenomenon that has been found in diverse settings for several decades (Jeannerod and Pacherie 2004, 123). The authors go on to consider a general account of how these and other resources are involved in ascribing a naked intention, namely the central monitoring theory (CMT), and conclude that CMT should be improved by adding on some some elements of Jeannerod's ST.

8.1.3 Central Monitoring Theory

The starting point for the discussion is provided by CMT. As this theory plays an important in various discussions that are relevant to our interests here, let me take a moment to sketch its basic features. On this theory, decisions to perform actions cause models of those actions to be sent forward to a comparator, which compares sensory and proprioceptive inputs with predictions of changes to the body and the environment stemming from the action to be carried out (Sperry 1950, Wolpert et al 1995).

How does this help us explain how we determine whether the actions we represent are our own or those of an observed agent? The proposal is to combine CMT with the concept of shared representations – in other words, to add on a simulationist component. Although CMT does not predict the phenomenon of shared representations, it does not rule them out either. The basic idea is that in action observation, no decision to perform an action has been made, and therefore no model of the action would be present in the comparator. The default ascription, then, in the absence of endogenous action-related signals from an internal model, is an ascription of the represented intention to someone else. Hence, the presence or absence endogenous action-related signals from such an internal model could distinguish between action performance and action understanding. The shared representation constitutes a naked intention, then, which is indifferent between action performance and action understanding and therefore cannot on its own be ascribed to oneself or to another person.

The notion of a naked intention, also referred to as an agent-neutral representation, is relevant to the discussion of MNs, since it presents a way of theorizing the contribution of MNs to action understanding in combination with other resources. After all, what we are looking for in that context is a way to understand how MNs could play an active and substantial role in action understanding, although they do not seem sufficient to account for all aspects of action understanding – such as contextual information and the ascriptive component (recognizing the representation of an intention as belonging to or referring to someone else). Since Jeannerod and Pacherie show us a way how the motor dimension, and perhaps also a goal, can be represented independently of an agent-representation, their theory gives us a good theoretical framework for the MN discussion. Thus, we will take it up again in chapter 7, when we return, armed with an expanded concept of simulation, to the discussion of MNs. For now, I would like to continue with the presentation of their theoretical account and discuss a type of pathological case that is relevant to explaining action understanding, namely schizophrenia.

Schizophrenia is a relevant pathological case insofar as some symptoms of schizophrenia, such as the belief that one controls or influences others' behavior or thoughts, or vice versa, can be regarded as involving either over-self-ascription or under-self-ascription of intentions or actions. In using this theoretical framework to explain some aspects of schizophrenia Jeannerod and Pacherie, are following Frick (Frith 2005, Frith, C. D., Blakemore, S.-J., & Wolpert, D. M. 2000a and 2000b) and Daprati et al. (1997), who both hypothesize that schizophrenic patients fail to monitor their action-related endogenous signals.

Experimental support for this hypothesis comes from experiments in which participants perform a movement and also see a hand on a video screen performing the same or a similar movement, and have to determine whether the hand they see is their own. Schizophrenics, as it turns out, are more prone to attribution errors than normal people (Daprati et al., 1997). But, interestingly, they are especially prone to over-attribute actions to themselves. And it is not clear that the explanation Frick offers in terms of CMT is well-suited to account for this fact. After all, in cases where they are not really performing an action, there should be no active internal action model from which action-related signals would come. Hence, there is no reason to expect them ever to over-self-attribute.

If under-self-attribution results from a failure to register action-related endogenous signals from an internal model, the complementary explanation of over-self-attribution would have to assert something along the lines of a retroactive activation of internal models. For CMT, this would be an ad hoc move, and it would also have the drawback that internal models would then be active in action observation as well as action performance, and could therefore no longer serve to distinguish between the two. Instead, other contextual cues would have to be decisive. But it is the line that Jeannerod proposes, and it follows naturally from his simulation theory of action.

Another problem with Frith's account, which Pacherie and Jeannerod (2004) think ST can improve upon, is that schizophrenics have no special problems with fine motor skills that would seem to draw upon coordination of visual information and proprioception. For example, Fourneret et al. (2001) ran an experiment in which subjects had the task of drawing a straight line, which they could not see, while the trajectory of the line they were drawing was presented to them on a computer screen. Unbeknownst to the participants, a bias was introduced to the system such that they in fact had to draw a curved line in order to make a straight line appear. Hence, there was a conflict between visual information and proprioception, (but no question of identifying the agent of the action, as in the Daprati experiment mentioned about one page ago). The result that is of interest to us here is that schizophrenics did not significantly differ from other participants, which bodes ill for Frith's hypothesis. In this case, the schizophrenic participants had no special problems monitoring actions by comparing proprioceptive and visual information with an internal model, and correcting on the fly to maximize the match between proprioceptive and visual information, on the one hand, and endogenous signals from the internal model, on the other. Moreover, most were not impaired in their ability to report verbally on the strategy they had employed, i.e. compensating for the bias.

Does Jeannerod's ST fare better? It is important to note that ST is not intended as a rival to CMT, but as an addition to it. It takes its starting point from a purported shortcoming of CMT. The shortcoming stems from the fact that CMT was not designed to account for covert actions – neither observed actions nor imagined actions (nor action planning). The combination just discussed (CMT plus simulation theory of action) offers a way to account for our differentiation between observed and executed actions. But note that the default ascription in the absence of an internal model is to another agent. And therein lies the problem – if we accept that our own imagined actions utilize the same shared representations as observed actions of others, sometimes the activation of the shared representation should be interpreted not as representing someone else's action but as representing our own imagined action (or action planning). In such cases, the default ascription would be false. It seems, then, that we need additional resources for distinguishing our own covert actions from action observation.

ST helps in these respects. It proposes that we simulate all covert actions, i.e. when we observe others and interpret their movements as constituting actions, when we imagine or recall our own actions, etc. Thus, it follows that over-self-attribution just as much a possibility as under-self-attribution. Why? Because, according to ST, there is always a simulation of an action occurring whenever an action is represented, irrespective of whether that action is being performed, observed, recalled, etc. This is a good theoretical outcome, since the empirical data also reveals that both phenomena occur.

ST also helps to integrate the phenomenon of verbal hallucination in schizophrenic patients. The idea is that hallucinators produce inner speech but fail to attribute the production of that verbal thinking to themselves, and thus experience the inner speech as arising from some other person speaking. In fact there is plenty of evidence of a connection between hallucination of voices and inner speech in schizophrenic patients. For example, the hallucinations occur more frequently when the patients are given addressed, given instructions, or otherwise led to produce inner speech, and often correspond to what would be appropriate responses to verbal stimuli (Chadwick and Birchwood 1994). Moreover, during hallucinations, the patients show muscular activity in the laryngeal muscles (Gould, 1949, David, 1994), which are of course involved in real speech production.

CMT does not have a good explanation of this phenomenon, since there is no vocal utterance during inner speech, and therefore no internal model of the action from which endogenous action-related signals (such as the expectation of auditory input as a result of one's own speech) would come. ST, on the other hand, predicts that much of the brain activity during verbal thinking should be the same as when hearing others speak. Hence, ST would predict such misattributions if and when there are problems with the other resources that are normally involved in the attribution to oneself or to someone else, or in differentiating between inner speech and vocalized utterances produced by oneself.

So, the next question ahs to be: what other areas are these? Using PET scans, McGuire et al. (1996) found that, when asked to produce inner speech or to imagine someone else speaking, patients predisposed to such hallucinations have reduced activity in regions in the frontal lobe associated with the production of speech. It has also been found that brain metabolism increases in the primary auditory cortex during such hallucinations. Hence, during hallucinations, the auditory system behaves as if it were processing the speech of an external speaker. The explanation that Pacherie and Jeannerod give on the basis of ST is that the brain simulates speech during silent speech, and that the areas in the frontal lobe that are less active in hallucinators normally have the function of informing speech perception areas of imminent language output so that the verbal stimuli is recognized as self-generated.

One point I would want to add is that the simulation of speech production during silent speech includes the auditory input that would normally accompany one's own vocalized utterances. Although Pacherie and Jeannerod do not say this explicitly, it appears to be a necessary feature of their account. After all, why else would the primary auditory cortex be active during silent speech? Still, the point is not trivial, and thus warrants being mentioned, since it would also be theoretically possible that the primary auditory cortex is simulating hearing someone else speak as opposed to hearing oneself speak. There is no way to discriminate between these two explanations by looking only at the primary auditory cortex. But, in view of the fact that there are also overlaps (such as the activation of the laryngeal muslces) between hallucinations and one's own production of vocalized utterances, it is sensible to follow Pacherie and Jeannerod and regarding this activation in the primary auditory cortex as constituting part of a simulation of one's one speech production.

What we have just been discussing in fact constitutes one special case of the general theoretical problem of accounting for how neutral representations are attributed to oneself or to others. Jeannerod speaks of "who detectors", which is a functional term for whatever achieves this aspect of action understanding. Physiologically, what one would have to look for is differences in brain activation across the various ways of representing the same action (i.e. performing it, recalling it, observing it, etc.). Without going into much detail about this, I will just mention some of the differences that have been found within the motor system. If you get bored easily by neurophysiological details, skip the following paragraph.

In the sensory-motor area (SMA), activation is more rostral (=frontal) during imagined and observed movements than during performance (Grafton et al1996, Gerardin et al., 2000). In the premotor cortex, Broadman area 6 is more active during covert actions than overt performance of actions (Binofski et al., 1999). In the cerebellum, imagined or observed actions produce more posterior activation (Lotze et al., 1999). In the primary motor cortex, I am not aware of any specific differences in location of activation, but there is a difference in intensity, which could serve the same function. Specifically, activation during motor imagery is about 30% of the level during performance (Gerardin et al., 2000). Outside the motor system, the STS is active while observing biological motion (Allison et al., 2000), and could play a role in signaling that there is another agent who is up to something.

So, what does all this mean? The most important point for our purposes is that simulation theory of action helps us to articulate what simulative processes in general, and MNs as a specific example, could contribute to action understanding, as well as what they do not contribute. Specifically, the concept of agent-neutral representations seems tailor-made to account for the functional role of simulations in general, and for MNs in particular. We have noted with respect to ST and MN research that the attribution of a representation of an action or an intention to someone else appears to be distinct from simulation of the action (in functional terms) or activation of areas associated with performing the action/having the intention (in physiological terms).

In summary, although our primary interest is with simulation in the narrower sense of social cognition, i.e. simulating others and not necessarily simulating one's own past or future actions, ST of action provides indirect support for ST of action understanding and for robust interpretations of MNs. Aside from the specific like simulation theories of concepts, this work embeds ST within a broader context, revealing that the overlap between action and thinking about action that ST claims is in fact consistent with a trend in neuroscience. This should impress us as a demonstration that ST increases theoretical coherence and unity among various approaches to understanding the mind. We will say a bit more later on about this trend, which can be labeled embodied cognition.

As a final point on simulation theory of action, it is of historical interest that this work began in the mid-1980s – around the time that ST was proposed in the folk psychology debate. There does not appear to have been a direct influence either way, so I would interpret this temporal coincidence as a sign of the broader trend toward embodied cognition that ties together the numerous simulation theories under discussion here.

8.2 Metacognition

Another, related application of the term simulation in psychology and ethology that has recently attracted the attention of philosophers of mind is the work on metacognition. The term "metacognition" refers to cognitive monitoring and control of first-order cognitive processes. In other words, it refers to cognitive processes that target other cognitive processes as opposed to events or properties in the world. The targeted cognitive processes include judging the adequacy of a particular response, correcting that response, evaluating one's ability to carry out a particular task[85], evaluating the ease or difficulty of learning some new information or of recalling some previously learned information (Proust 2006, 18-19)[86]. Ease of learning has probably been the most extensively researched area: subjective assessments of ease of learning prior to learning to tasks, judgment of learning during and after learning tasks, and feeling of knowing – i.e. judgment about whether a currently nonrecallable item will be remembered in a subsequent test (Nelson 1992).

Their directedness to cognitive processes immediately suggests an analogy to folk psychology, which, after all, is also supposed to be some kind of cognitive apparatus for tracking the cognitive processes of others. The difference in the case of metacognition, at least in the work done so far, is that metacognitive processes are directed toward one's own cognitive processes. Despite this difference, metacognition research is of interest to us because it presents empirical data and theoretical resources that should be taken into consideration and can be fruitful it theorizing about the closely related case of folk psychology. After all, in our pursuit of understanding mental states and mental concepts, we certainly want to compare and contrast first- and third-person cases, and assure a certain amount of symmetry[87]. I will say more about this in a moment, but for starters let me note that there are efforts to conceive of the kind of representation involved in metacognition without invoking metarepresentation, which is a central goal of ST. Indeed, as I will explain, the term simulation plays a prominent role in metacognition. Being applied to first-person cases here, the term "simulation" is extended in the same way as in simulation theories of concepts or of

[85] This includes cases of perceptual assessment of the feasibility of a task, , but also assessment of one's ability to perform mental tasks, such as finding a solution to a logistical problem.

[86] Nelson et al. 1992 and Koriat 2000 and Koriat 1993 are influential theoretical presentations, Proust 2006 & 2007 present the state of the art in the theoretical discussion.

[87] Although not perfect symmetry, because of the phenomenon of privileged access.

action à la Jeannerod. This is no accident; we will see that metacognition researchers draw upon the same ideas from dynamic system theory, and that they share a concept of simulation derived from that theoretical background.

But let's back up for a moment and start out by sketching the influential theoretical account of metacognition given in Nelson et al. (1992), since it provides a simple sort of "common denominator" picture of the *structure* of metacognition, which, as far as I can tell, everyone involved in the discussion basically accepts. The structure postulated by Nelson et al. is based upon three principles:

1) A distinction between object-level and meta-level. With respect to learning tasks, for example, this implies that learning procedures are employed on an object-level, and that these processes are represented on a meta-level.
2) The meta-level contains a simulation of the object-level. This is a closer specification of what principle 1 refers to as representation at the meta-level. Note the ambiguous relationship between metarepresentation and simulation. Insofar as we are talking about representation on a meta-level, it seems we must be talking about metarepresentation. But, as we know, simulation can be conceived as an alternative to metarepresentation. We will come back to this in a moment.
3) There are two kinds of relation that obtain between the two levels: control and monitoring. In monitoring, events on the object level are registered, observed, tracked or something along these lines (for the moment leave open the choice between metarepresentation and simulation) on the meta-level. Here, the information flows from object-level to meta-level, so we have the metacognitive equivalent of a mind-to-world direction of fit. In control, events at the object-level are influenced by signals issuing from the meta-level.

The most obvious way to conceive of the representational vehicles involved in metacognition is to appeal to metarepresentation, i.e. to stipulate that the meta-level metarepresents the object-level. There is some developmental support for this theoretical approach, namely in the developmental synchrony of numerous abilities that appear to depend on metarepresentation. So we find that until around the age of four years, children are poor at false belief tests and also at linking their knowledge with events in the past that gave rise to that knowledge (Perner & Ruffman, 1995). For example, if they answer a question with a lucky guess, they claim that they knew they answer just as often as if they really had known it, had visual access to the contents of a drawer or been informed by a reliable source (Perner 1991). And, even if they did acquire the answer by reliable means, they tend to forget how, and not accurately state whether they were told or discovered the answer on their own.

But there is also comparative data that speaks against conceiving of metacognition as a kind of metarepresentation. Specifically, metacognitive abilities have also been documented in non-human species such as monkeys, dolphins and birds (Smith et al., 2003), who show no other signs of having the ability to represent mental states. For example, Smith (in press) conducted a comparative study involving rhesus monkeys and humans, in which participants were confronted with a visual discrimination task. The interesting thing was that they had the option of declining the task in favor of an easier test that was less rewarded. This exit-option was therefore a wise alternative in cases where the visual discrimination task was especially difficult or when for any other reason participants were not confident in their ability to complete it. Hence, deciding whether to take the task or to exit required the participants to make a metacognitive assessment of their own epistemic security with respect to the visual discrimination task. The result was that the monkeys behaved quite similarly to the human participants.

The reason why these findings from comparative primatology present a problem for the view that metacognition is a species of metarepresentation is that other indicators of metarepresentational abilities are absent in these same species that apparently have metacognitive abilities. In particular, their ToM abilities are at best quite limited – no non-human species succeed at false belief tests. Their failure at false belief tests speaks against attributing a mastery of mental concepts to them. And if they have no mental concepts, then they cannot be metarepresenting their own mental states as mental states, since this would require categorizing those mental states with a concept that they lack. Joelle Proust (2006, 2007) therefore concludes that metacognition does not involve the ability to metarepresent (one’s own) mental states.

Instead, Proust regards metacognition as a simpler form of representation than metarepresentation. She needs, then, to give an account of what kind of representation is involved in metacognition if it is not metarepresentation. What she is looking for is a way of linking the meta-level with the object-level that does not include the content of representations at the object-level within the content of representations at the meta-level. The meta-level has to track events at the object-level that correlate with success at the task at hand. So, for example, if the task is to memorize some visual stimulus, say a particular arrangement of dots on a screen, there may be a neural activation that is typical of instances in which the arrangement has successfully been learned. This pattern will be identified by the monitoring mechanism at the meta-level. But the actual visual percept need not be represented at the meta-level. In fact, the specific content of the visual percept can be considered irrelevant at the meta-level. What is being represented at the meta-level is not the visual percept but a feature of the cognitive process being monitored, namely its chance of success. The way that feature is represented at the meta-level may be constant across various kinds of monitored processes, and would therefore not correlate with any particular feature of the world. In other words, the semantic properties of the object-level representation need not be carried over into the representation at the meta-level. As Proust writes, "The control structure establishes a link between observed feedback and new command, and reciprocally; but it does not need to use the current contents of the corresponding epistemic states to secure this link." (Proust 2006, 26).

As phylogenetic support of this general picture, she points out an interesting asymmetry: there are species that have metacognitive abilities but lack ToM abilities, whereas there are no species that have ToM abilities but lack metacognitive abilities. Proust thinks this indicates that metacognition is an older adaptation, and that metarepresentation and ToM are built upon it. For the moment, my aim is not to assess the plausibility of this claim[88], but to show the role that simulation plays in Proust's theory of metacognition[89]. So, the idea is that metacognition enables organisms to simulate an action covertly in order to evaluate its chances of success (Proust 2006, 20). This would be an improvement upon always having to carry out an action in order to evaluate it, and would thus save time and other resources and help to avoid danger. Simulation as a good way of testing out options covertly.

Of course the alternative is also available to stick to the concept of metarepresentation in theorizing about metacognition, if one is willing to qualify the concept of metarepresentation accordingly. One could admit that non-linguistic animals can metarepresent and then give some other explanation of why they cannot apply this skill in certain kinds of case, such as mindreading. But there are also theoretical reasons to pursue the alternative strategy of maintaining that metacognition is functionally distinct from metarepresentation. Joelle Proust (2007) argues that metacognition exhibits certain properties that are absent from metarepresentation (causal contiguity, epistemic transparency and procedural reflexivity), while metarepresentation also exhibits certain properties that are absent from metacognition (open-ended recursivity and inferential promiscuity). She proposes the following features of metacognitive engagement:

[88] Thank goodness. It does seem plausible, but that is true of a lot of just-so stories…

[89] Her theoretical framework builds upon Smith et al. 1995 and Koriat 1993.

1) predictive or retrodictive
2) a self-directed evaluative process
3) has a normative and motivational function
4) not explicable in first- order terms (because it relates not to the world but to the subjective feeling of knowing)
5) not explicable in second-order terms (this would make it meta-represenstational, but I do not need to entertain the concept of the activity that is being exercised in order to evaluate it, since I can rely on feedback derived from expected sensory input or a feeling of its running smoothly)

In short, there is support for the view that metacognition constitutes a perspective upon one's own mental processes that does not require metarepresentational resources such as mental concepts. For this reason, it is highly interesting in the context of ToM research, in particular with respect to efforts to theoretically account for the fledgling ToM skills that children display in the years prior to full mastery of mental concepts (e.g. as revealed in false belief tasks around 3.5). The issue is also relevant to our understanding of adult ToM skills, since numerous authors from very different theoretical backgrounds (Gallagher 2005, Perner 1991, Gordon 1986) urge that mature adult mindreading does not replace but supplement the mechanisms employed by young children. It may prove fruitful to regard the non-conceptual structure of metacognition as a model for (basic) mindreading skills in children and in non-linguistic creatures?

In effect, all of these simulationist theories can be regarded as representative of the trend toward embodied cognition that has been increasingly popular since the 1980s. Let me now conclude this chapter by saying a bit about this embodied cognition movement to which I have had occasion to refer from time to time throughout the dissertation.

8.3 Embodied Cognition

8.3.1Embodied Representations

Although there are various positions within the embodied cognition movement and, as we have seen, various versions of ST, ST can be said broadly to fit into the overall embodied cognition movement. The primary reason why I say this is that the understanding of representation that is common among propoonents of embodied cognition is closely related to notions of representation that are appropriate to ST and, in particular, to notions of ST that interpret MN research robustly. The basic idea underlying embodied representations is that they are not couched in an amodal symbolic code that exists in addition to sensorimotor perceptual representations. Rather, representations are distributed throughout the sensorimotor system, the entire body and the environment. Since research has focussed on perceptual representations, let's take as an example the perceptual representation of an object. According to embodiment theories, such a representaiton includes the object itself as well as the possibilities to interact with that object. So, representing a glass means being inclined to have certain expectations about future perceptions and being aware of certain options. One would, for example, expect to hear a loud crash if one swept one's ahnd across the area where the glass is located. One would also be aware of the option of having a receptacle to pour a liquid into if one had the desire to do so. The glass does not have to be additionally represented aside from these expectations and awarenesses, since the glass itself is already there. There is no need for an additional (representation of a) glass in one's head. Instead one can "use the world as its own best representation", as Brooks puts it (1993). One also sometimes calls such representations "virtual models", since the models do not actually exist anywhere I particualr (e.g. in the head), but the organism acts as though they did.

This program is similar to ST insofar as ST claims that I use my own possibilites to act instead of additional folk psychological resources in order to understand others' actions. Concepts of actions are in this sense, embodied in my own possibiliites for action. Embodied cognition approaches are useful and interesting for ST insofar as they deal with how such various elements of representations (e.g. objects in the world, the motor system, modal sensory systems) can interface in cognitive processes. I would like to illustrate this with a couple of examples of work on embodied cognition.

8.3.2 Metaphors and Conceptual Thought

There is some very interesting recent empirical work by Chen-Bo Zhong and colleagues that supports the general idea that states of the body are invovl. Zhong's research has targeted the interaction between social emotions and perceived states of the body, such as bodily temperature and cleanliness. As for body temperature: they have found that people who have been excluded from a social activity are more inclined to report feeling cold, to assess the room temperature as being lower than it is, and to desire warm food and beverages. This suggests an influence of social emotions upon perceived bodily temperature. Conversely, room temperature and the temperature of food and beverages that one received also influence one's judgments of how friendly other people are (Zhong, C. & Leonardelli, G., 2008), suggesting that perceived bodily temperature influence social emotions. As for cleanliness: people who are asked to recall morally questionable acts that they have committed or to copy down a fictional first-person narrative about committing misdeeds are more inclined to prefer antiseptic hand wipes than to other objects, and are more likely to fill choose sanitary words like "soap" and "wash" to complete fragments like S_ _ P and W_ _ H, rather than neutral words like "step" and "wish" (Zhong, C. & Liljenquist, K. 2006). This suggests an influence of self-assessment of moral integrity upon self-assessment of bodily cleanliness. The influence appears to work the other way as well: people who are allowed to wash themselves are less likely to volunteer to perform altruistic acts.

This work supports some general conclusions that I have been homing in on. Firstly, it supports the view that there is a combination of various kinds of representations in conceptual thought. In this case, whatever resources are involved in assessing temperature and personal cleanliness are shown to interact with judgments about morality and with social cognition and social emotions, such as whether another person is friendly or one's subjective feeling of being either excluded from or integrated in a group. Secondly – this is a closely related point – it supports the view that higher-level cognition, in particular social cognition, is embodied in the sense that it draws upon one's own bodily states and processes.

To relate this latter point to the concept of simulation, one could say that this line of research suggests that we simulate cleaning or getting dirty when assessing the morality of certain actions, or that we simulate being cold or warm in assessing another person's friendliness or our own relationship to a person or a group. I do not intend to sell this as a proof of ST, but to bring out the continuity between this work and ST and to show that embodied cognition approaches employ – either implicitly or explicitly – a concept of simulation that is broader than but includes the concept of simulation we find in ST.

8.3.3 Common code theory

Wolfgang Prinz has developed a “common-coding approach” to understanding the relationship between perception and action. The basic idea is that action planning/performance and action understanding/perception rely on a common (or commensurate) code. Prinz’s approach is of interest for several reasons – because it emphasizes events in the world as the common denominator between perception and action; because it supports and helps to clarify the notion that concepts integrate various kinds of representations, including but not limited to perceptual and motor representations; and because it helps to contextualize ST within the broader trend toward embodied cognition.

Let me start out by saying what the common-coding approach is not. Prinz contrasts common-code theory with separate-coding approaches, which, he says, have dominated psychology for decades or perhaps centuries (Prinz 2005, 141). Separate-coding approached assume that actions are coded as muscular movements and perceptions are coded as patterns of stimulation of sensory organs. It is easy to see, given this starting point, why information about muscular movements and information about the sensory organs do not appear to be commensurable. In other words, there needs to be some sort of *translation* of perceptual information into muscular information in order for perception to help guide ongoing actions, or for muscular movements to be selected in order to bring about desired events in the world (which are perceptible and thus represented in the perceptual code as opposed to the action code).

In contrast to that, Prinz thinks that perception and action are coded in the same format, thus rendering them commensurable and making translation superfluous. To see how this is supposed to work, consider Prinz’s definition of action: “any meaningful segment of an organism’s intercourse with its environment” (Prinz 1990, 167). This formulation implies that actions are represented in terms of their effects in the world, which are of course perceptible. Hence, it avoids the dichotomy between action and perception from the start by identifying actions with perceptible events.

Prinz regards common-coding theory as building upon the theory of voluntary action proposed by R. Lotze and W. James. The basic idea of this theory, as interpreted by Prinz, is that: “…the representations of the intended goal states have the power of generating the action directly, that is, without the need for any further volitional activity” (Prinz 2005, 142). That is, when I represent the world as being a certain a way, in the absence of inhibitive factors, I automatically perform whatever movements bring about that state of the world. So, to take an example from James, if I am given the task of pressing the left of two keys in response to stimulus A and the right key in response to stimulus B, and I see stimulus A, I only have to think of the appropriate event (pressing the left key) 'and – "presto!" – it happens by istelf'.[90] Of course this is merely an appeal to phenomenological plausibility, but it is not intended at this stage to be anything more. At the very least, it clarifies the intuition at the basis of the theory: that in choosing and planning actions, we think about perceptible events, not about muscular movements. As Prinz puts it:

> Any representation of an event of which we learn that it goes along with, or follows from, a particular action will afterward have the power to elicit the action that produces the event (Prinz 2005, 143).

Of course, there is still a difference between the perceptual representations and action representations by virtue of their direction of fit. Action representations specifiy how they world should be, whereas perceptual representations specify how it is. Prinz acknowledges this: "The distinction between percept and act codes is not anchored in the distinction between the body and its environment, but rather between event representation and event effectuation." (Prinz, W. 1990, 172)

Prinz claims that the common-coding is well-placed to offer explanations of a few different phenomena. One point that Prinz emphasizes, and which his theory is designed to account for, is that intentional processes (e.g. the selection of goals) influence perception in such a way as to optimize the extraction of information relevant to the intentional state. This influence is facilitated by the fact that motor and perceptual representations are couched in the same code. Another explanandum that the common-coding theory addresses is the phenomenon of imitation. To see why common-coding theory is well-placed to explain imitative behavior, consider the following: imitation, conceived as "performing an act after and by virtue of seeing it done by someone else," depends on *similarity* between a perceived act and a performed act (Prinz 2005, 142). The common representation format asserted by the common coding theory asserts just such a similarity between representations of perceived actions and representations of performed actions. So, just as thinking of an event leads me to produce that event in the absence of inhibition, so does seeing someone else produce that event.

To help clarify Prinz's position and to give an idea of what testable empirical difference it might make, I will take a moment to describe an experimental paradigm with which Prinz seeks to garner support for common-coding theory. It involves two kinds of

[90] From Williams James' *Principles of Psychology* (1890), as quoted by Prinz 1990, 171.

tasks, namely *matching tasks* and *mapping tasks*. An example of a matching task would be the following: subjects must respond to visual presentations of letters [e.g., A,B,C] by producing the corresponding vocal utterance. Mapping tasks differ in that the link between the stimulus and the response is arbitrary. Prinz's assertion is that the separate-coding theory does not predict that matching tasks should be any easier (measured in terms of speed and accuracy), whereas the common-coding view does. Prinz (1990) reports results that bear out his prediction.

To sum up, the common-coding approach is attractive for a few reasons. *First*, the common denominator of perceptual representations and action representations in common-coding theory is events in the world, which accords well with my commitment to retain the spirit of Gordon's ST as much as possible and thus to try to avoid invoking mental concepts, introspection and analogical inference for as long as possible. Insofar as mental concepts do need to be added, they include perceptual and motor representations associated with a desired event. This is a programmatic point. *Secondly*, since the basic idea of the common-coding approach is that perceptual representation and action representation are commensurable, or couched in the same format, his framework offers a way of combining the perceptual representations that are at the focus of simulationist theories of concepts with motor representations. This is attractive because the interpretation of MNs that I have been supporting requires that MNs be enriched by other resources, including but not limited to perceptual representations, contextual information, the presence or absence of efference copies and proprioception.

8.4 General Remarks

The aim in this chapter has been threefold. *Firstly*, I have tried to spell out more precisely what it could mean to simulate one's own actions, and what role that could play in simulating and understanding others' actions. *Secondly*, I have tried to clarify the idea of agent parameters or who detectors, which helps to characterize the additional resources that must come into play in order to distinguish between (simulations of) one's own actions and (simulations of) others's actions. *Thirdly*, a further aim in this chapter has been to show how ST, in particular when modified to incorporate simulations of one's own actions, fits into the broader paradigm of embodied cognition. This helps us to understand the contribution of ST to the discussion on social cognition in seberal ways. It gives us a corpus of empirical work that could potentially be invoked as support for ST, assuming that further theoretical work on

ST is pursued with a mind to incorporating this corpus of empirical work. Also, like the rest of the work reviewed here, it gives us some suggestions about the limitations of simulation in social cognition or conceptual thought generally.

Chapter 9:
What Has Been and What Shall Come to Pass:
A recap of the main points raised so far and some thoughts about where to go from here

9.1 Recap of what's happened so far

In the two opening chapters, I introduced the main ideas underlying the folk psychology discussion and attempted to demonstrate why this discussion has been of interest to philosophers. I attempted to show there that the interest in folk psychology reflected a natural way to think about the mind-body problem subsquent to the linguistic turn: to leave ontological issues aside and investigate the way in which mental concepts function in explanations of behavior, how they relate to other concepts, whether they are reducible, eliminable, etc. I argued that a further shift began in the 1980s and continues today: the assumption that our everyday psychological competence is best accounted for by postulating a folk psychology featuring mental concepts and psychological generalizations is increasingly viewed as problematic. One important cause of this shift was the appearance of an alternative account of folk psychology, namely simulation theory (ST), at which point the hitherto unquestioned account came to be known as theory theory (TT). With respect to philosophy of mind, a primary attraction of ST is that it undermines the threat of eliminativism by denying that folk psychological concepts and laws ever really played a role in everyday social cognition at all.

Apart from embedding TT in a disursive trajectory stretching from Ryle and Sellars to Lewis, Fodor, Dennett & co., I also discussed the historical context of ST. Although I acknowledged the connection between ST and the hermeneutic tradition, I suggested that ST also has historical roots in the tradition of psychophysical parallelism stretching form Fechner to Mach and Hering and on to Piaget. Reviewing this tradition, I point out that ST can benefit by engaging with the parallelist view of the mind-body relation, according to which mental and physical events are token-identical but are distinct by virtue of the epistemic access that we have to them, mental events being accessible from the first-person perspective and physical events from the third-person perspective. This fits well with ST insofar as ST is committed to the view that our understanding of minds (especially our own) is not exhausted by the kind of theoretical knowledge that a functionalist (or theory theory) account of mental states would allow us to attain by drawing inferences from observable behavior. For ST, there has to be some more direct way of understanding our own mental lives and also the mental lives of other people. The parallelist tradition offers a solution in the distinction between first-

and third-person perspectives. A benefit gained by opting for such a position is that it elegantly accounts for the privilegeed access that we take ourselves intuitively to have to our own minds.

On the other hand, a burden of this position is that it forces one either to accept introspection as the means of ascribing mental states from the first-person perspective or to develop some other alternative. I suggest that the version of parallelism developed by Piaget is a good starting point for developing an alternative, since it conceives of privileged access not as identifying occurrent beliefs and desires, for which we would need introspection, but as a means of predicting actions that follow from constellations of beliefs and desires, in particular in light of the fact that we expect others to behave rationally. Specifically, we can use our own intuitions about what is rational in lieu of a theory of rationality. Piaget's conception eneables us to abandon TT's idea that people in everyday life employ nomological psychological generalizations to predict others' behavior. We are still left with the theoretical task of accounting for how we represent, ascribe (to ourselves and to others) and conceptualize the beliefs and desires that figure in these predictions.

In the third and fourth chapters, I analyzed various versions of TT and ST. In the fifth chapter, I reviewed empirical work intended to discriminate between ST and TT, and also looked at alternative views. My conclusion was that there is no uniform distinction to be made between the two competing theories; rather, the various versions represent a broad range of ideas, many of which may prove useful in ongoing empirical research. I point out that this is to be expected if one takes seriously the analogy to scientific practice invited in particular by TT but also by ST, since simulations are generally developed, utilized and interpreted in combination with various theoretical elements in scientific practice.

My strategy in the second half of the dissertation was to start out with what I take to be the basic insight of ST – that in making sense of others' behavior *we undergo the same procedures that we would undergo if we ourselves were deciding upon, planning or executing an action in the same circumstances* – and to ask how this idea can be refined to fit with and contribute to ongoing empirical research.

Chapter 6 is devoted to interpretations of mirror neurons (MNs), since the discovery of MNs constitutes support for ST, insofar as ST would predict that resources for action (e.g. the motor system) would be used in action understanding. But since action understanding appears to require a more abstract kind of representation than motor representation (since one action can be carried out with different movements and different actions can be carried out with one and the same movement in different contexts) and to incorporate contextual information, the role of mirror neurons is likely to be contingent upon their integration with other areas (e.g. STS, SMC). I argue that the incorporation of these additional ingredients can be accomplished by expanding the concept of simulation to include simulations of *one's own* past or imagined perceptual experiences along the lines of embodied (a.k.a. perceptual or simulationist) theories of concepts (Barsalou 1999, 2003, Prinz 2002, Lakoff and Gallese 2004).

This approach, which is the theme of chapter 7, offers a way to show how action concepts (as well as other concepts) could be represented in a more abstract way than the motor system or perceptual systems alone could, but without introducing amodal symbols in the sense of a language of thought. Extending this account to mental concepts, as would be necessary at least for some forms of action understanding, I draw upon Lakoff's theory of concepts, according to which one of the parameters of an action concept is the agent parameter. An agent parameter could be realized in the brain by neural activation present during performance but not observation of an action (e.g. SMA, primary motor cortex). MNs can function, therefore, as component parts of mental concepts in action understanding. Research on action representation, including theoretical accounts of the so-called who-system and embodied cognition generally, are the focus of chapter 8. This material also offers interesting perspectives upon how agent-neutral representations are combined with other factors to distinguish between one's own actions and those of others, suggesting a unified account of mental concepts for first- and third-person ascriptions while still leaving room for privileged access to one's own intentions at least in most cases.

In chapter 8, I also discuss the idea of metacognition. Theoretical work on metacognition incorporates a notion of simulation that is first-person directed (i.e. we simulate our own acitons in planning them, selecting them, and assessing their chances of success. This notion is especially interesting in light of the need that I have diagnosed for an account of first-person access to one's mental states that is distinct from full-blown introspection and that does not presuppose full conceptual mastery. The further issue, of course, is whether this first-person access can be of use in folk psychology, i.e. whether it can be applied via simulation to others.

9.2 Where to go from here

Joelle Proust, one of the main theoreticians of metacogntion, does not think that metacognition can be involved in third-person interpretation (for which she uses the term "mind-reading"). She points out that "one cannot compare in an engaged way (metacognitively informed) the reafferences imputed to another subject with the norm *that* subject acquired via his previous experiences." (Proust, 2007, 25) This objection is of a kind with a common objection to simulation theory, namely: in order to determine whether the other person is relevantly similar to us or to imaginatively take their perspective and correctly set the parameters for a simulation, we need background knowledge and third-person generalities that exploit mental concepts. Hence a conceptually minimalist mode of access such as metacognition might offer cannot suffice.

This is a reasonable concern, which we should take seriously. I think that it can be addressed, though, if we abandon the presupposition that everyday social cognition should be primarily conceived as third-person interpretation, and consider that it is more plausible to conceive of social cognition as primarily interactive, or collaborative. If we take up this point and shift our focus from third-person mindreading to interactive, collaborative mindreading, then Proust's reservations about applying metacognition to mindreading carry less weight.

This move is motivated also by an important difference between children and apes, namely that children are highly motivated to engage in joint attention, joint action, and other forms of social interaction (Tomasello 1999, 2008). The emerging consensus in developmental and comparative psychology is that non-human primates are able to understand conspecifics' states of mind (e.g. intentions, beliefs, desires) to some extent, but that *they are less motivated than humans to do so*. They do so only to the extent that others' minds are immediately relevant to the production of concrete events in the environment, whereas humans are interested in others' states of mind for the sake of sharing attention and other states of mind and collaborating in joint activities.

The theoretical problem here is to account for young children's (i.e. in the years before 3.5) understanding of others' attentional states, goals, etc. in collaborative contexts without invoking mental concepts. Note that this mirrors the theoretical problem of accounting for metacognition in non-human species. I think it may prove fruitful to combine these two problems, and seek an account of the semantics of metacognition that is applicable to the empirical data on joint attention and joint activity in young children. One the one hand, research on metacognition could profit by the drawing upon distinctions entrenched in this body of research. It would be interesting, for example, to learn more about how young chlidren fare at non-verbal tests of metacognition in various kinds of collaborative contexts, depending on whether the motivation is individual, shared or altruistic. On the other hand, metacognition is well-suited, in particular by virtue of its motivational component, to offer a model of the kind of representational structure that children draw upon in collaborative contexts. Specifically, children may draw upon metacognitive skills in collaborative contexts before they have the ability to metarepresent their own or others' mental states. Practice in such collaborative contexts could also contribute to the development of metarepresentational skills.

In working out the semantic structure and epistemological status of whatever proto-mental concepts are at work in metacogntion, it seems reasonable to start out from success semantics, which gives an account of beliefs that is applicable to non-linguistic creatures. The basic idea (Ramsey 1927) is that beliefs are functions from desires to actions. Jose Bermudez (2003) builds upon this idea, specifying beliefs in terms of utility conditions, which, in turn, are understood in terms of the satisfaction of desires, i.e. the conditions under which the behavior ceases. Although success semantics admits of a great deal of indeterminacy in ascribing beliefs, this is not necessarily a drawback. In fact, this epistemological indeterminacy may be seen as adequately reflecting the content-indeterminacy of the vehicles of representation involved in social cognition in pre-linguistic children and even in adults in many cases. The theoretical work of identifying such limitations of success semantics and developing resources to improve upon them can also have a heuristic function in empirical research. The central role played by desires in success semantics is especially attractive in light of the emerging consensus in developmental and comparative psychology (Tomasello 2008) that non-human primates are less motivated than humans to engage in joint activity for the sake of sharing attention and cooperating. They may understand others' intentional states to some extent insofar as these are instrumental in producing effects in the environment that interest them, but not insofar as they are instrumental to the fulfillment of joint desires or others' desires, thus limiting their ability to ascribe beliefs in the sense of success semantics.

This rough account of metacognition is intended to show how ST can incorporate a deflationary account of privileged access to one's own mental states that could be applied to also to third-person intterpretation. In order for this extension to third-person cases to work, though, there has to be some mechanism to assure that our own minds are similar to those of others. The discussion in chapter 5 about the analogy to science can help us to see how this could work.

Again, Ian Hacking's treatment of experimental practice within the debate on scientific realism suggests an interesting way to cash out on the analogy between folk psychology and science (now on an epistemological level). Hacking asserts that the history of practical application and refinement of scientific equipment and models (and, I would add, simulations) gives us reason to think that they are linked up with the world in a systematic way that has epistemic import. Analogously, I would argue the following: the distinctions that we make in folk psychological discourse (between propositional attitudes such as believing, wanting, and fearing) guide us during development in imitating others and thereby learning from them and becoming more similar to them. One result, of course, is that we become increasingly reliable simulation systems. But the flipside of this is even more interesting: since in cognitive development we use folk psychological notions (such as goals, strategies, beliefs and desires, and attentional states) to guide our simulations (either in actually imitating others or in trying to understand what they are up to), we become increasingly similar to other people *as they appear to us when we employ folk psychology (understood as a combination of theoretical elements and simulation on the basis of our habits/dispositions)*. So other people will be able to effectively employ the same sort of folk psychological resources upon us. In short, the role of folk psychology in cognitive development ensures that it will be an effective way of predicting and explaining others' behavior.

To put this another way, the idea is that folk psychology is a self-validating system. Employing it upon others during develpment has the effect that it can be employed upon us ever more effectively. There is a similarity between this idea and Kusch's (sec. 5.2.3) idea that folk psychology is self-referential. I can agree with Kusch that mental states are therefore at least partially socially constructed. But the argument that I would give is that this social construction of mental states in fact engages causally with cognitive development in such a way that mental concepts are linked up with the real processes going on in people's heads and causing their behavior. Mental concepts can be said to refer to these processes (whether they should be construed as entities, processes, events or something else is a further question). In short, there is therefore also a substantial realist component that is missing from Kusch's account and thtat I think is highly important to include.

Thus, there is good reason to believe that the distinctions referred to in folk psychology reflect real structures among the psychological causes of our behavior. Insofar as the way in which we employ mental concepts in giving psychological explanations is intertwined with simulations, (i.e. habits and dispositions) during development, the latter could be an irreplaceable resource for scientific psychology. This means that simulation theory (surprisingly) gives us reason to think that mentalist description is not only compatible with scientific psychology, but is indeed potentially useful *as such* to scientific psychology. So, if you like Hacking's entity realism, you should also consider becoming a realist about mental states.

A central point that I have been driving at is that folk psychology works not because it is an accurate theory, but because its use has the effect that one winds up being similar to how one seems to others, and can therefore also be interpreted by others who use it. It seems plausible that this ability does not primarily involve postulating theoretical entities, since the initial grasp of intentions comes from having had a particular kind of experience of an external object, e.g. of wanting it, which has contributed to shaping our habits and dispositions in such a way that they are similar to those of other people. If we use our own dispositions as a predictive basis in folk psychology, we may simply be monitoring our own states in the manner of meta-cognition, which may be lacking central features of meta-representation, such as inferential promiscuity and open-ended recursivity[91]. In other words, our exploitation of our own habits and dispositions in folk psychology does not need to be explained by invoking meta-representation of our own representational states, and thus does not need to involve the application of mental concepts. So the belief or desire that plays a role in folk psychology does not refer to a private object in someone's mind or brain, but to an experience of the world. In another sense, though, it is private: only we can metacognitively monitor our own thought processes. This provides an elegant account of our intuition that we have *privileged access* to our minds without having to incorporate introspection.

Of course, we also learn to use theoretical elements to enhance our ability to predict people's behavior, but there are two points to make about this. *Firstly*, these theoretical elements often are nothing other than information about the world, about social narratives, about people's typical behavior and social roles, etc. In other words, the theoretical elements are not hidden entities inside their minds or brains. *Secondly*, when explicitly mental concepts such as representations of others' belief states are involved in predicting their behavior (in secondary cases such as those involving false beliefs), the usefulness of these mental concepts hinges upon their being combined with our own habits and dispositions. We do not need to explain the success of explanations that invoke mental concepts by arguing that the mental concepts must successfully refer; the habits and dispositions that are also involved in the explanation are already linked up with the habits and dispositions of the person whose behavior is being explained. Hence, accounting for our folk psychological practices and intuitions does not require us to postulate mysterious objects that would violate the principle of *causal closure* of the physical world.

[91] For a thorough discussion of the distinction between metacognition and metarepresentaton, see Proust 2007. Metacognition and Metarepresentation: Is a self-directed theory of mind a precondition for metacognition? *Synthese* 159 (2), 271-295.

As for the *intentionality* of mental states, the point is that the practice of ascribing intentions and then imitating them and otherwise orienting our behavior accordingly has the effect that we can usefully be regarded as intentional agents. What this amounts to in naturalistic terms is still best accounted for by an information-theoretic approach à la Dretske[92]/Millikan[93]. The basic idea here is that mechanisms in the brain (at least having to do with perception, and perhaps with cognition more generally) evolved to produce brain states that have the biological function of co-varying reliably with states of affairs in the world, and in this sense to give information about, or to represent, those states of affairs in the world. This account is good as far as it goes. But it ignores culture, and also runs into the problem that there is a complex and potentially quite long causal chain leading up to any brain state. How do we decide where to stop tracing that chain backward and pick out one link in the chain as the intentional content? This problem can be dealt with by introducing the influence of culture into the account, which is exactly what my proposal invites us to do. The basic idea is simply that convention determines at what point we stop re-tracing the causal steps that led to a mental state and pick out a thing, a process or an event as the content. And since the very cognitive mechanisms that go into forming our mental states are shaped by our initiation into these conventions, the conventions reliably reflect the features of the environment that are relevant to the cognitive processes in question.

As for *semantic opacity*: the problem is that you cannot know the content of someone's intentional state just by knowing a physical object they are referring to; you have to know under what description they know it, which you can in principle only do if you know all their beliefs. This is considered to be a challenge for empirical psychology. My point would be that we can limit this problem by reflecting that we can interpret people's intentional states insofar as people are similar to us. So we can use the principles of charity and assume that others will attend to the same perceptual information as we attend to, and that the inferences they derive from their beliefs and desires will be such as to seem coherent to us. The challenge is to know what background knowledge they have and what goals they are pursuing. In the context of joint activities, this is often fairly obvious, even explicit. And we can narrow down the hypotheses when there is ambiguity by observation, by trial and error, and even by asking questions. In scientific psychology, I would suggest, we can and do use the same sorts of procedures to narrow down hypotheses about people's mental states. It is no insurmountable stumbling block to naturalism if we accept that there is not likely to be a

[92] Dretske F. (1981). *Knowledge and the flow of information*, Blackwell: Oxford.

[93] Millikan R. (1984). *Language, Thought, and Other Biological Categories: New Foundations for Realism*, Cambridge: MIT Press.

single, definitive account of the mental states relevant to a given inquiry; it is enough to narrow down hypotheses and give reasons for finding one or the other of them most plausible, just like in other sciences. The only difference is that in psychology simulation can fruitfully be employed as well.

Abstract

What is Folk Psychology and Who Cares?
The folk psychology debate from the perspective of philosophy of mind

(English)

The dissertation presents and critically discusses the debate between two rival theories of folk psychology – theory theory (TT) and simulation theory (ST). After a contextualization of the folk psychology discussion within the history of philosophy, I analyze the leading versions of both theories. My conclusion is that there is no uniform distinction to be made between the two competing theories; rather, the various versions represent a broad range of ideas, many of which may prove useful in ongoing empirical research. In the second half of the dissertation, I discuss recent work on social cognition in neuroscience that should taken into consideration in further developing a theory of folk psychology. A long chapter is devoted to interpretations of mirror neuron (MN), since this work constitutes support for ST, insofar as ST would predict that resources for action (e.g. the motor system) would be used in action understanding. But since action understanding appears to require a more abstract kind of representation than motor representation (since one action can be carried out with different movements and different actions can be carried out with one and the same movement in different contexts) and to incorporate contextual information, the role of mirror neurons is likely to be contingent upon their integration with other areas (e.g. STS, SMC) I also discuss theories of concepts that make it possible to integrate such elements within a simulatinist framework.

(Deutsch)

Die dissertation präsentiert und diskutiert die Debatte zwischen zwei Theorien der Alltagspsychologie – Theorie-theorie (TT) und Simulationstheorie (ST). Nach einer historischen Kontextualiserung der beiden Theorien, werden die wichtigsten Versionen der zwei Theorien vorgestellt und kritisiert. Meine Schlussfolgerung ist, dass keine einheitliche Unterscheidung zwischen den zwei Theorien vorgenommen werden kann. Sie stellen vielmehr ein breites Spektrum an theoretischen Optionen dar, die bei der weiteren Entwicklung einer Theorie über die Alltagspsychologie berücksichtigt werden sollten. In der zweiten Hälfte der Dissertation diskutiere ich neuere Arbeiten in den Neurowissenschaften, die für diese Thematik relevant sind. Ein längeres Kapitel ist den sogenannten Spiegelneuronen gewidmet, weil diese Forschungsrichtung einen wichtigen Beleg für die ST darstellt, insofern als die ST vorhersagen würde, dass Ressourcen, die für die eigenen Handlungen gebraucht werden, auch für die Deutung der Handlungen anderer Menschen benutzt werden. Das Verstehen von Handlungen verlangt aber eine abstraktere Repräsentation als die Spiegelneuronen leisten könnten, da Handlungen durch motorische Bewegungen unterbestimmt sind (da eine Handlung mit verschiedenen Bewegungen ausgeführt werden könnte und verschiede Handlungen mit ein und derselben Bewegung ausgeführt werden können). Daher ist der Beitrag der Spiegeleuronen zum Verstehen von Handlungen wohl im Zusammenhang mit anderen Bereichen (z.B. STS, SMC) zu vermuten. Ich diskutiere auch Begriffstheorien, die es ermöglichen, solche Elemente innerhalb eines simulationstheoretischen Rahmens zu integrieren.

Bibliography

Adams, Fr. R. (2001). Empathy, neural imaging and the theory versus simulation debate. Mind and Language, 16, 368-392.

Adolphs R. (2003) Cognitive neuroscience of human social behaviour. *Nat Rev Neurosci*, 4(3):165-178.

Adolphs R, Tranel D, Damasio D, Damasio A. (1994). Impaired Recognition of Emotion in facial expression following bilateral damage to the amygdala, Nature 37: 1111-1117. Adolphs, Ralph, 1995, "Fear and the human amygdala", *Journal of Neuroscience* 15: 5879-5891.

Adolphs R., Damasio H, Tranel D, Cooper G, and Damasio A. (2000). A role for somatosensory cortices in the visual recognition of emotion as revealed by three-dimensional lesion mapping. J. *Neurosci*, 20, 2683-2690.

Allison T et al. (2000). Social Perception from the Visual Cues. Role of the STS region. *Trends in Cognitive Sciences* 4: 267-78.

Altshuler E et al. (2000). Social perception from visual cues: role of the STS region. In *Trends in Cognitive Sciences* 4: 267-278.

(1997). Person see, person do: human cortical electrophysiological correlates of monkey see monkey do cell. Society of Neuroscience Abstracts 719.17.

Apperly I. (2008). Beyond Simulation-Theory and Theory-Theory: Why social cognitive neuroscience should use its own concepts to study "theory of mind". Cognition, vol 107 (1): 266-283.

Appleton M, Reddy V. 1996). "Teaching three-year-olds to pass false belief tests: a controversial approach", *Social Development* 5, 275-291.

Arbib M. (2005). The Mirror System Hypothesis. Linking Language to Theory of Mind, 2005, retrieved 2006-02-17

Astington J, Gopnik A. (1991) Developing understanding of desire and intention. In Whiten (1991).

Avis J. and Harris P. (1991), "Belief-desire reasoning amon Baka children: evidence for a universal conception of mind," Child Development, 62, 460-7.

Baillargeon R. (1987). Object permanence in 3.5 and 4.5 month old infants. Developmental Psychology 23: 655-64.

Baron-Cohen S. (1995). Mindblindness: An Essay on Autism and Theory of Mind. Cambridge: MIT Press.

(1991). Precursors to a theory of mind: understanding attention in other. In: Whiten (1991).

Baron-Cohen S, Joliffe S, Mortimore T, Robterson M. (1997). Another advacned test of theory of mind: evidence from very high functioning adults with autis or Asperger syndrome. Journal of Child Psychology and Psychiatry 165: 813-822.

Baron-Cohen S, Ring H, Wheelwright S, Bullmore E Brammer M, Simons A, Williams S. (1999). Social intelligence in the normal and autistic brain: an F.M.R.I. study. *European Journal of Neuroscience* 11, 1891-1898.

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

Baron-Cohen S, Leslie A, and Frith U. (1986). "Mechanical, behavioral an intentional understanding of picture stories in autistic children", British Journal of Developmental Psychology, 2, 113-25.
(1985). Does the autistic child have a 'theory of mind'? *Cognition* 21: 37-46.

Baron-Cohen S, Tager-Flusberg H, Cohen D. (2000). *Understanding other minds: Perspectives from developmental cognitive neuroscience*. (2nd Ed.) Oxford: OUP.

Baron-Cohen S, Tager-Flusberg H, Mitchell P (eds.) (1993). *Understanding Other Minds: Perspectives from Autism*. Oxford: OUP.

Barsalou L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22, 577-609.

(1993). Flexibility, Structure, and Linguistic Vagary in Concepts: Manifestations of a Compositional System of Perceptual Symbols. In Collins A, Gathercole A, Conway M, Morris P (eds.). *Theories of Memory*. Hillsdale: Lawrence Erlbaum Associates.

(1992). Frames, concepts, and conceptual fields. In A. Lehrer & E. F. Kittay (Eds.), *Frames, fields, and contrasts: New essays in lexical and semantic organization*, pp. 21-74. Hillsdale, New York: Erlbaum.

(1989). Intra-concept similarity and its implications for inter-concept similarity. In Vosniadou S and Ortony S (eds.). *Similarity and Analogical Reasoning*. NY: Cambridge Univ. Press.

(1987). The instability of graded structure: Implications for the nature of concepts. In U. Neisser (Ed.), *Concepts and conceptual development: ecological and intellectual factors in categorization*, pp. 101-140. Cambridge, UK: Cambridge Univ. Press.

Barsalou L., Santos A, Simmon W, Wilson Chr. D. (2008). Language and simulation in con- ceptual processing. In M. de Vega & A. M. Glenberg & A. C. Graesser (Eds.), *Symbols and embodiment: Debates on meaning and cognition*, pp. 245-283. New York: Oxford University Press.

Barsalou L, Solomon K, and Wu L. (1999). Perceptual simulations in conceptual tasks. in: Hiraga, Masako (ed.). International Comparative Literature Association : *Cultural, psychological and typological issues in cognitive linguistics* : selected papers of the bi-annual ICLA meeting in Albuquerque, July 1995 / ed. by Masako Hiraga ... . - Amsterdam: Benjamins 1999 . - 338 S. . - 1-55619-867-1. - (Amsterdam studies in the theory and history of linguistic science : Series 4, Current issues in linguistic theory ; 152)

Bartsch K, Wellman H. (1995). *Children talk about the mind*. New York: OUP.

(1989) Young children's attribution of action to beliefs and desires. *Child Development*, 60, 946-64.

Berkeley G (1710). *A treatise concerning the principles of human knowledge*. Dancy J (ed.). (1998). Oxford: OUP.

Berlin B, Kay O. (1969). *Basic Color Terms. Their Universality and Evolution*. Berkeley: Univ. Cal. Press.

Bermudez J. (2003) *Thinking without Words*. Oxford: Oxford Univ. Press.

Binet, A. [1905], *L'Ame et le Corps*, Ernest Flammarion: Paris, 221-232.

F. Binkofski, G. Buccino, S. Posse, R. J. Seitz, G. Rizzolatti and H. J. Freund, A fronto-parietal circuit for object manipulation in man. *Eur. J. Neurosci.* 11 (1999), pp. 3276–3286.

Bird C, castelli F, Malik O, Frith U and Husain M. (2004). The impact of extensive medial frontal lobe damage on 'theory of mind' and cognition. *Brain* (127) No. 4, 914-928.

Blackburn S. (1995). Theory, observation, and drama. In Davies M and Stone T. (1995a).

Blair R.J.R., Mitchell D.G.V., Peschardt K.S., Colledge E., Leonard R.A., Shine J.H., Murray L.K. and Perrett D.I. (2004). "Reduced sensitivity to others' fearful expressions in psychopathic individuals," *Personality and Individual Differences* 37: 1111-1122.

Blair R.J.R., Mitchell D.G.V., Richell R.A., Kelly S., Leonard R.A., Newman C., Scott S.K. (2002). "Turning a deaf ear to fear: Impaired recognition of vocal affect in psychopathic individuals," *Journal of Abnormal Psychology* 111: 682-686.

Bonda E, Petrides M, Ostry D and Evans A. (1996). Specific involvement of human parietal systems and the amygdale in the perception of biological motion. *Journal of Neuroscience* 15: 3737-3744.

Bos E, Jeannerod M. (2002). Sense of body and sense of action both contribute to self-recognition
Cognition vol. 85, Issue 2,
Pages 177-187

Brooks R.A (1993). The engineering of physical grounding. In: *Proceedings of the fifteenth annual conference of the cognitive science society*, Lawrence Erlbaum, Hillsdale, NJ (1993), pp. 153–154.
(1991). Intelligence without representation. Artificial Intelligence, 47, 139-159.

Bruner J. (1990). Acts of Meaning. Cambridge MA: Harvard Univ. Press.

Buccino G et al. (2001). Action observation asctivates premotor and parietal areas in a somatotopic manner: an fMRI study. In The European Journal of Neuroscience 13: 400-404.

Butterworth G. (1991). The ontogeny and phylogeny of joint visual attention. In: Whiten (1991).

Camerer, C., Loewenstein, G., and Weber, M., (1989). “The curse of knowledge in economic settings: an experimental analysis,” Journal of Political Economy 97: 1232-1254.
Carruthers P. (2009) How we know our own minds: The relationship between mindreading and metacognition. Behavioral and brain Sciences, vol 32 (2): 121-138.

(1996) Autism as mind-blindness: an elaboration and partial defence. In Carruthers and Smith (1996).

Carey S. (1985). Conceptual Change in Childhood. Cambridge: MIT.

Carlson A, Moses L, Hix H. (1998). The role of inhibitory processes in children’s difficulty with false belief. Child Development 69: 672:691.

Carruthers P, Smith P, (ed). (1996). Theories of Theories of Mind. Cambridge: Cambridge University Press.

Chadwick, P. and Birchwood, M. 1994: The omnipotence of voices. A cognitive
approach to auditory hallucinations. British Journal of Psychiatry, 164, 190–201.

Churchland P. (1981). Eliminative Materialism and Propositional Attitudes, Journal of Philosophy, 78: 67-90.

Cochlin S et al (1999). Observation and execution of movement: similarities demonstrated by quantified electroencephalography. In The European Journal of Neuroscience 11: 1839-1842.

Collingwood R. (1946). The Idea of History, Oxford: Clarendon.

Colunga, E. & Smith, L.B. (2003). The emergence of abstract ideas: Evidence from networks and babies. In L. Saitta (Ed.): Philosophical Transactions by the Royal Society B. Theme Issue: The Abstraction Paths: From Experience to Concept. 358 (1435), 1205-1214.

Csibra G. (2008), Action Mirroring and action understanding: an alternative account. In Haggard et al (2008).

(2005). Mirror neurons and action understanding. Is simulation involved?
http://www.interdisciplines.org/mirror.

Currie G, Ravenscroft I. (2004). Recreative Minds: Imagination in philosophy and psychology. Oxford: Clarendon Press.

Damasio, Antonio, 1999, The Feeling of What Happens, New York: Harcourt Brace.
(1989). Time-locked multiregional retroactivation: a systems-level proposal for the neural substrates of recall and recognition. Cognition 33: 25-62.

E. Daprati, N. Franck, N. Georgieff, J. Proust, E. Pacherie, J. Dalery and M. Jeannerod, Looking for the agent. An investigation into consciousness of action and self-consciousness in schizophrenic patients, Cognition 65 (1997), pp. 71–86.

Davidson D. (2001). Subjective, intersubjective, objective, Oxford: Clarendon Press.

(1970). Mental Events. In Foster and Swanson (eds.). Experience and Theory, London: Duckworth (Reprinted in Davidson 2001).

David, A.S. 1994: The neuropsychological origin of auditory hallucinations. In A.S. David and J.C. Cutting (eds), The Neuropsychology of Schizophrenia. Lawrence Erlbaum, Hove, pp. 269–313.

Davies, M. and Stone, T. (eds.) (1995a). Folk Psychology: The Theory of Mind Debate. Cambridge, MA: Blackwell.

(1995b). Mental Simulation: Evaluations and Applications. Cambridge, MA: Blackwell.
Decety, J. 2002 Is there such thing as a functional equivalence between imagined, observed, and executed action? In The imitative mind: development, evolution, and brain bases (ed. A. N. Meltzoff & W. Prinz), pp. 291–310. Cambridge University Press.

Decety J, Jackson P (2004). The functional architecture of human empathy. Behavioral and Cognitive Neuroscience Reviews, Vol. 3, No. 2, 71-100.
Decety, M. Jeannerod, D. Durozard and G. Baverel, Central activation of autonomic effectors during mental simulation of motor actions. J. Physiol. 461 (1993), pp. 549–563.

DeLoache J. (1989). The development of representation in young children. In: Reese W (Ed). Advances in child development and behavior, vol. 22: 1-39. New York: Academic Press.

(1987) Rapid change in the symbolic functioning of very young children. Science 238: 1556-1557.

Dennett D. (1978). Beliefs about beliefs. Commentary on Premack and Woodruff (1978) Behavioral and Brain Sciences 1 (4): 568.

(1981). Three Kinds of Folk Psychology, Reduction, Time and Reality Healey, R. (ed.): 37-61 New York: Cambridge University Press.

(1981) Making sense of ourselves. Philosophical Topics, 12 (1).

(1987). The Intentional Stance, MIT Press, Cambridge.

Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. Experimental Brain Research, 91, 176-180.

Dokic J. and Proust J. (ed.)., 2002, Simulation and Knowledge of Action. Advances in Consciousness Research 45, John Benjamins

Dretske F. (1981). Knowledge and the flow of information, Blackwell: Oxford.

Evans G. (1982), The Varieties of Reference, edited by John McDowell, Oxford: Oxford University Press.

(1981). Reply to Crispin Wright: Semantic Theory and Tacit Knowledge. In Holtzman S, Leich C. (eds.). Wittgenstein: To Follow a Rule. London: Routledge.

Feldman, J.E. (2006). From molucule to methaphor. A neural theory of language. Cambridge, Mass.: MIT Press.

Feldman, J.E. & Narayanan, S. (2004). Embodied meaning in a neural theory of language. Brain and Language, 89, 385-392.

Fadiga L, Fogassi L, Pavesi G, Rizzolatti G. (1995) 'Motor facilitation' during action observation. Amagnetic stimulation study'. In Journal of Neurophysiology 73: 2608-2611.

Ferrari, P. F., Gallese, V., Rizzolatti, G., & Fogassi, L. (2003). Mirror neurons responding to the observation of ingestive and communicative mouth actions in the monkey ventral premotor cortex. European Journal of Neuroscience, 17, 1703-1714

Flavell J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. American Psychologist, v34 n10 p906-11 Oct 1979.

Flavell J, Everett B, Croft K, and Flavell E. (1981). Young children's knowledge about visual perception: further evidence for the level1-level2 distinction. Developmental Psychology 17: 99-103.

Flavell J, Flavell E, Green F and Moses L. (1990). Young children's understanding of fact beliefs versus value beliefs. Child Development 61: 915-28.

Flavell J, Green F, Flavell E (1986). Devlopment of knowledge about the appearance-reality distinction. Monographs of the Society for Research in Child Development, 51 (serial no. 212).

Fodor J. (2000). There are no recognitional concepts – not even RED. In Fodor J. In Critical Condition: Polemical Essays on Cognitive Science and Philosophy of Mind. Cambridge: MIT Press.

(1998). Concepts: Where Cognitive Science Went Wrong, (The 1996 John Locke Lectures), Oxford University Press

(1994). The Elm and the Expert, Mentalese and its Semantics, (The 1993 Jean Nicod Lectures), MIT Press.

(1991). A Modal Argument for Narrow Content. Journal of Philosophy 88: 5-25.

(1983). The Modularity of Mind. Cambridge: MIT Press.

(1975). The Language of Thought. Cambridge: Harvard Universtiy Press.

Fodor J, Lepore E. (1992). Holism: A shopper's guide. Oxford: Basil Blackwell.

T Fogassi, L. & Gallese, V. (2002). The neural correlates of action understanding in non-human primates. In M. I. Stamenov & V. Gallese (Eds.), Mirror Neurons and the Evolution of Brain and Language (pp. 13-35). Amsterdam: John Benjamins Publ.he Modularity of Mind, Cambridge: MIT Press.

Follesdal, D, and Quine, D. (2008), Quine in Dialogue, Harvard University Press: Harvard, 2008.
Fourneret, P.; Franck, N.; Slachevsky, A.; Jeannerod, M. (2001) Self-monitoring in schizophrenia revisited. Neuroreport 8 May 2001 - Volume 12 - Issue 6 - pp 1203-1208
Cognitive Neuroscience And Neuropsychology

Firth R. (1980). "Epistemic merit, intrinsic and instrumental". Presidential Address to the Eastern Division of the American Philosophical Association.

Frith, C. D. (2005). The self in action: Lessons from delusions of control. Consciousness and Cognition,
14(4), 752–770.

Frith, C. D., Blakemore, S.-J., & Wolpert, D. M. (2000a). Abnormalities in the awareness and control of
action. Philosophical Transactions of the Royal Society of London B, 355, 1771–1788.

Frith, C. D., Blakemore, S.-J., & Wolpert, D. M. (2000b). Explaining the symptoms of schizophrenia: Abnormalities in the awareness of action. Brain Research Reviews, 31, 357–363.

C.D. Frith and U. Frith, The neural basis of mentalizing, Neuron 50 (4) (2006), pp. 531–534.

Frith U. (1989). Autism: Explaining the Enigma, Oxford: Blackwell.

Gallagher S (2006). Logical and phenomenological arguments against simulation theory. In: Hutto and Ratcliffe (2006).

(2005). How the body shapes the mind. Clarendon Press: Oxford.

Gallese V. (2005) Embodied simulation: from neurons to phenomenal experience. Phenomenology and the Cognitive Sciences 4: 23-48.

Gallese, V. (2004). Intentional attunement. The mirror neuron system and its role in interpersonal relations. http://www.interdisciplines.org/mirror

(2003). A neuroscientific grasp of concepts: from control to representation. Phil. Trans. R. Soc. Lond. B (2003) 358, 1231-1240.

Gallese V. 2001. The ‘shared manifold’ hypothesis: from mirror neurons to empathy, Journal of Consciousness Studies, 8, 33-50

Gallese, G.; Fadiga, L.; Fogassi, L.; Rizzolatti (1996), "Action recognition in the premotor cortex", Brain 119 (2): 593–609, http://brain.oxfordjournals.org/cgi/content/abstract/119/2/593

Gallese V., and Goldman A. 1998. Mirror neurons and the simulation theory of mind-reading, Trends in Cognitive Sciences 2, 493-501

Gallese, V. & Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in conceptual knowledge. Cognitive Neuropsychology, 22, 455-479.

Gallese V and Metzinger T. (2003). Motor ontology: the representational reality of goals, actions and selves. Philosophical Psychology, vol. 16, No.3.

N. Georgieff and M. Jeannerod, Beyond consciousness of external reality: A “Who” system for consciousness of action and self-consciousness. Conscious. Cogn. 7 (1998), pp. 465–472.

Gérardin, E., Sirigu, A., Lehe´ricy, S., Poline, J-B., Gaymard, B., Marsault, C., Agid, Y. and Le Bihan, D. 2000: Partially overlapping neural networks for real and imagined hand movements. Cerebral Cortex, 10, 1093–1104.

Gergely G, Csibra G. (in press) Syvia’s Recipe: The role of imitation and pedagogy in the transmission of cultural knowledge, in: N.J. Enfield, S.C. Leveneson (ed), Roots of Human Sociality: Culture, Cognition, and Human Interaction: Berg Publishers.

Gergely G , Nadasdy G, Csibra G and Biro S. (1995). Taking the intentional stance at 12 months of age. Cognition 56: 165-93.

Gastaut H, Bert J. (1954). ‘EEG changes during cinematographic presentation.’ In Electroencephalography and Clinical Neurophysiology 6: 433-444.

Glenberg, A.M. and Kaschak, M.P. Grounding language in action.
Psychon. Bull. Rev. (in press)

. (2003). The body's contribution to language. In B. Ross (Ed.), The Psychology of Learning and Motivation, V43 (pp. 93-126). New York: Academic Press.

Goldman A. (2006). Simulating Minds: The philosophy, psychology, and neuroscience of mindreading. Oxford: Oxford University Press.
(2005). Mirror Neurons, Social Understanding and Social Cognition. http://www.interdisciplines.org/mirror.

(1995) Interpretation Psychologized. In Stone T and Davies M (eds.) Folk Psychology: The Theory of Mind Debate. Oxford: Blackwell.

(1993). The Psychology of Folk Psychology. In Goldman A (ed.). (1991). Readings in Philosophy and Cognitive Science. Cambridge MA: MIT Press, 347-380.
Goodman N (1976). Languages of Art: An approach to a theory of symbols. Indianapolis: Hackett.

Gopnik A. (1996). Theores and modules: creation myths, developmental realities, and Neurath's boat. In Carruthers and Smith (1996).

Gopnik A, Astington J. (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, Meltzoff A. (1996). Words, Thoughts and Theories. Cambridge: MIT.

(1993). The role of imitation in understanding persons and in developing a theory of mind. In Baron-Cohen S, Tager-Flusberg H, Mitchell P (eds.) (1993). Understanding Other Minds: Perspectives from Autism. Oxford: OUP.

Gopnik A. and Wellman H. (1995). Why the child's theory of mind really is a theory. In Davies and Stone (1995).

Gordon R. (2005). Mirroring phenomena as a natural kind. http://www.interdisciplines.org/mirror

(2004). Intentional Agents Like Myself. In S. Hurley & N. Chater (Eds.), Perspectives on Imitation: From Cognitive Neuroscience to Social Science, Cambridge, MA: MIT Press, in press.

(2003). "Folk Psychology as Mental Simulation”, The Stanford Encyclopedia of Philosophy (Winter 2003 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2003/entries/davidson/>

(1996). Radical Simulationism. In Carruthers and Smith (1996).

(1995). Simulation without introspection or inference from me to you. In Stone T and Davies M (eds.) Mental Simulation: Evaluations and Applications. Oxford: Blackwell.

(1986). Folk psychology as simulation. Mind and Language, 1, 158-171.

(1992a). The simulation theory: objections and misconceptions. Mind and Language, 7, 11-34.

(1992b). Reply to Stich and Nichols. Mind and Language, 7, 85-97.

(1992c). Reply to Perner and Howes. Mind and Language,7, 98-103.

(1987) The Structure of Emotions: Investigations in Cognitive Philosophy. London: Cambridge University Press.

Grezes, J. and Decety, J. (2001) Functional anatomy of execution,

Vol.7 No.2 February 2003
S. T. Grafton, M. A. Arbib, L. Fadiga and G. Rizzolatti (1996). Localization of grasp representations in humans by positron emission tomography. 2. Observation compared with imagination. Exp. Brain Res. 112 (1996), pp. 103–111

J. Grezes, C.D. Frith and R.E. Passingham, Inferring false beliefs from the actions of oneself and others: An fMRI study, Neuroimage 21 (2004), pp. 744–750
J. Grezes, C.D. Frith and R.E. Passingham, Brain mechanisms for inferring deceit in the actions of others, The Journal of Neuroscience 24 (24) (2004), pp. 5500–5505

Gould, L.N. 1949: Auditory hallucinations in subvocal speech: objective study in a case of schizophrenia. Journal of Nervous and Mental Diseases, 109, 418–427.

Hacking I. (1983). Representing and Intervening: introductory topics in the philosophy of science, New York: Free Press.

Hampton J. (1995). Testing the prototype theory of concepts. Journal of Memory and Language 34: 686-708.

(1991). The combination of prototype concepts. In Schwanenflugel P. (ed.). The Psychology of Word Meanings. Hillsdale: Lawrence Erlbaum Associates.
Happé F et al. (1996). 'Theory of Mind' in the brain. Evidence from a PET scan study of Asperger Syndrome. NeuroReport 8: 197-199.

Haggard P, Y. Rossetti and Kawato M. (Eds.). (2008). Sensorimotor Foundation of Higher Cognition: Attention and Performance XXII, Oxford: OUP.
Haller R. and Stadler F. [1988] (eds), Ernst Mach: Werk und Wirkung, Wien: Hölder-Pichler-Tempsky, 1988.

Hari R et al. (1998). Activation of human primary cortex during action observation: a neuromagnetic study. Proceedings of the Academy of Sciences of the USA 95: 15061-15065.

Harman G. (1978). Studying the chimpanzee's theory of mind. Commentary on Premack and Woodruff (1978) Behavioral and Brain Sciences 1 (4): 576.

Harris, P. (1991). The work of the imagination. In Whiten)1991).

Heal J. (2003). Mind, Reason, and Imagination: Selected Essays in Philosophy of Mind and Language. London: Cambridge University Press.

(1996), Simulation, theory and content. In Carruthers, Peter and Smith, Peter (eds.). (1996), pp. 75-91.

(1995). How to think about thinking. In Stone T and Davies M (eds.), Mental Simulation: Evaluations and Applications. Oxford: Blackwell.

(1994). Simulation versus theory-theory: what is at issue? Proceedings of the British Academy, 83, 129-144.

(1986). Replication and functionalism. In Butterfield J (ed.), Language, Mind and Logic (pp. 135-150) Cambridge: Cambridge University Press.

Heidelberger M (2004). Nature From Within. Gustav Theodor Fechner and his Psychophysical Worldview. Pittsburgh: University of Pittsburgh Press.

(2000). Der psychophysische Parallelismus: Von Fechner und Mach zu Davidson und wieder zurück. In Elemente moderner Wissenschaftstheorie. Zur Interaktion von Philosophie, Geschichte, und Theorie der Wissenschaften, edited by Friedrich Stadler. Vienna: Springer, 91-104.

(2000). Fechner und Mach zum Leib-Seele Problem," in Materialismus und Spiritualismus. Philosophie und Wissenschaften nach 1848, Andreas Arndt and Walter Jaeschke (Hr.), Hamburg: Meiner.

[1990]: 'Concepts of Self-Organization in the 19th Century', in Krohn, W., Küppers, G. and Nowotny, H. (eds), 1990, Selforganization: Portrait of a Scientific Revolution, Dordrecht: Kluwer 1990 (Sociology of the Sciences, Yearbook 1990, Bd. XIV), 170-180.

F. Heider and M. Simmel, An experimental study of apparent behavior. Am. J. Psychol. 57 (1944), pp. 243–249.

Heim I. (2002). File change semantics and the familiarity theory of definiteness. In Portner P, Partee B (eds.) Formal Semantics: the essential readings. Oxford: Blackwell.

Herbart, Johann Friedrich, 1850, Bd. 5-7: "Schriften zur Psychologie", in Sämmtliche Werke, G. Hartenstein (Hr.), Leipzig: Voss.

Hering, E. [1876], 'Zur Lehre von der Beziehung zwischen Leib und Seele. I. Ueber Fechner's psychophysisches Gesetz.' Sitzungsberichte der Akademie der Wissenschaften, Wien, 72 (Abt.3): pp. 310-348.

Hogrefe G, Wimme H, Perner J. (1986). Ignorance versus false belief: A developmental lag in attribution of epistemic states. Child Development 57: 567-582.

Hughes C. (1998). Executie functioning in pre-schoolers: link with theory of mind and verbal ability. British Journal of Developmental Psychology 7, 237-250.

Hurley S. (2005). The shared circuits model. How control, mirroring, and simulation can enable imitation and mind reading. http://www.interdisciplines.org/mirror.

Hurley S and Nudds M (eds.). (2006) Rational Animals? Oxford. OUP.

Hutto D and Ratcliffe M (2007) Folk Psychology Re-Assessed. Springer: Dordrecht.

Iacoboni M, Molnar-Szakacs I, Gallese V, Buccion G, Mazziotta J, Rizzolatti G. (2005). Grasping the intentions of othes with one's own mirrow neuron system. PLoS Biology vol. 3 (3) e79.

Iacoboni M, Mazziotta JC (2007). "Mirror neuron system: basic findings and clinical applications". Ann Neurol 62: 213.

Jacob P (2008). What do mirror neurons contribute to human social cognition? Mind and Language vol 23 (2): 190-223.

Jackson F. (1982). Epiphenomenal Qualia, Philosophical Quarterly 32, 1982: 127-136.

Jacob P. (2008) What do mirror neurons contribute to human social cognition? Mind and Language vol. 23 (2): 190-223.

Jacob, P., Jeannerod, M., 2005, „The motor theory of social cognition: a critique", http://jeannicod.ccsd.cnrs.fr/ijn_00000573

(2003). Ways of Seeing. Oxford: OUP.
Jäger, S. (ed.)[1988], Briefe von Wolfgang Köhler and Hans Geitel, 1907-1920, Passau Univ. Press: Passau, p. 27.

James W. (1952). Principles of Psychology, Chicago: Britannica Great Books.

Jeannerod, M.(2000) Neural simulation of action: A unifying mechanism for motor cogntition. NeuroImageVolume 14, Issue 1, July 2001, Pages S103-S109

(1995) Mental imagery in the motor cortex. Neuropsychologia 33, 1419–1432

(1994). The representing brain. Neural correlates of motor intention and imagery. Behav. Brain Sci. 17: 187–245.

Marc Jeannerod & Elisabeth Pacherie (2004). Agency, Simulation and Self-Identification. Mind and Language 19 (2):113-146

Kaschak, M.P. and Glenberg, A.M. (2000) Constructing meaning: the role of affordances and grammatical constructions in sentence comprehension. J. Mem. Lang. 43, 508–529

Keil F. (1989). Concepts, Kinds and Cognitive Development. Cambridge: MIT.

Keysers C, Gazzola V. (2007). Integrating simulation and theory: from self to social cognition. Trends in Cognitive Sciences 11: pp. 153–157.

(2006), Towards a unifying neural theory of social cognition, Progress in Brain Research
Kilner J, Frith C (2007) Action observation: inferring intentions without mirror neurons. Current Biology (18) No. 1 R32.

Keysers, C., Wickers, B., Gazzola, V., Anton, J-L., Fogassi, L., and Gallese, V. (2004) A Touching Sight: SII/PV Activation during the Observation and Experience of Touch. Neuron: Vol. 42, April 22, 1-20.

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: 846-848.

Kim J. (1998). Mind in a Physical World, MIT Press: Cambridge.

Kohler et al (2002). Hearing Sounds, understanding actions: action representation in mirror neurons. Science 297: 846-848.

Klein A, Schultz R and Cohen D. (2000). The need for a theory of mind in action. In Baron-Cohen S, Tager-Flusberg H, Cohen D (eds.) (2000).

Knobe J (2008) Folk Psychology: science and morals. In Hutto and Ratcliffe (eds.) (2008).

(2006). The concept of intentional action: a case study in the uses of folk psychology. Philosophical Studies 130: 203-231.

Kögler, Hans Herbert, and Karsten R. Stueber, 2000, Empathy and Agency: the problem of understanding in the human sciences, London: Westview Press.

Köhler, Wolfgang, [1910] ‚Akustische Untersuchungen II,' ZfP, 58, pp. 98, 102

Koriat, A.: 2000, The Feeling of Knowing: some metatheoretical Implications for Consciousness and Control. Consciousness and Cognition, 9, 149-171.

Koriat, A.:1993, How do we know that we know ? The accessibility model of the feeling of knowing. Psychological Review, 100, 609-639

Kripke S. (1980). Naming and Necessity. Cambridge MA: Harvard Univ. Press.

(1979). A puzzle about belief. In Margalit A (ed.) Meaning and Use. Dordrecht: Reidel.

Kusch, M. (2008). Folk Psychology and Freedom of the Will. In Hutto and Ratcliffe (eds.) (2008).

(2006). Psychological Knowledge: a social history and philosophy, London: Routledge.

Langacker R. (1999) A view from cognitive lingusitics. In Behavioral and Brain Sciences, 22: 625.

Lakoff, G. (1994). What is a conceptual system? In W.F. Overton & D.S. Palermo (Eds.): The nature and ontogenesis of meaning. The Jean Piaget Symposium Series. pp. 41-90. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Lakoff G, Gallese V. (2004). The brain's concepts: the role of the sensory-motor system in conceptual knowledge Cognitive Neuropsychology, 2005, 22(3/4), 455-479.

Lakoff G, Johnson M. (1980). Metaphors we live by. Chicago: University of Chicago Press.

Lawrence, A.D:, and Calder, A.J., 2004, "Homologizing human emotions," in: E. Evans, and P. Cruse, (Eds.) Emotions, Evolution and Rationality (15-47), New York: Oxford University Press.

Lawrence S, Margolis E. (2000). Concepts and cognitive science. In: Lawrence S, Margolis E (eds.). (2000). Concepts: Core Readings. 2nd Print Cambridge, MA: MIT Press.

Leslie A. (1991). The Theory of Mind Impairment in Autism: Evidence for a Modular Mechanism of Development? In: Whiten (1991).

(1988). Some implications of pretense for mechanisms underlying the child's theory of mind. In: Astington J, Harris P and Olson D (Eds.). Developmental theories of mind, 19-46. New York: Cambridge Universtiy Press.

(1987) Pretense and Representation: the origins of 'Theory of Mind'. Psychological Review 94: 212-226.

Leslie A, German T. (1995) Knowledge and Ability in 'Theory of Mind': One-eyed Overview of the Debate. In Davies and Stone (1995b).

Leslie A, Roth D. (1993). What autism teaches us about metarepresentation. In. Baron-Cohen et al (1993).

Leslie A, Theiss L. (1992). Domain specificity in conceptual development: Neuropsychological evidence from autism. Cognition 43: 225-51.

Lewis D. (1972). Psychophysical and theoretical identifications, Australasian Journal of Philosophy 50:249-258.

Lichtermann L. (1991). Young Children's Understanding of Desires. Third year progress report (Unpublished manuscript).

Liu, D., Wellman, H.M., Tardif, T. & Sabbagh, M.A. (2008). A meta-analysis of false-belief understanding across cultures and languages. Developmental Psychology, 44, 523-531.

Lizskowski, U., Carpenter, M., Henning, A., Striano, T., & Tomasello, M. (2004). 12-month-olds point to share attention and interest. Developmental Science 7, 297–307.

Logothetis. (2008). What we can and what we cannot do with fMRI. Review. Nature 453, 869-878 (12 June 2008)

M. Lotze, P. Montoya, M. Erb, E. Hülsmann, H. Flor, U. Klose, N. Birbaumer and W. Grodd (1999). Activation of cortical and cerebellar motor areas during executed and imagined hand movements: An fMRI study. J. Cogn. Neurosci. 11: 491–501

Mach, E, [1906], Analyse der Empfindungen und das Verhältnis des Physisichen zum Psychischne, 5th ed., Jena: Fischer.

Margolis E, Laurence S. (2008). "Concepts", The Stanford Encyclopedia of Philosophy (Fall 2008 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2008/entries/concepts/>.

Matelli, M., Luppino, G. and Roizzolatti, G. (1985). "Patterns of cytochrome activity in the frontal agranular cortex of macaque monkey." Behav. Brain Res. (18): 125-137.

McGuire, P.K., Silbersweig, D.A., Murray, R.M., David, A.S., Frackowiak, R.S.J. and Frith, C.D. 1996: Functional anatomy of inner speech and auditory verbal imagery. Psychological Medicine, 26, 29–38.

Mead, George Herbert, 1934, Mind, Self and Society, Chicago: University of Chicago Press.
Meltzoff, Andrew N., 2005, „Imitation and other minds: The „Like Me Hypothesis“, in S. Hurley, N. Chater (Hr.), Perspectives on Imitation: From Neuroscience to Social Science Vol.2, Cambridge, MA: MIT, S. 55-77,

Meltzoff A. (1988a). Infant imitation after a one-week delay: Long-term memory for novel acts and multiple stimuli, Developmental Psychology 24: 470-476

(1988b). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children, Developmental Psychology 31: 838-850

(2005). Imitation and other minds: The “Like Me Hypothesis, in S. Hurley, N. Chater (ed.), Perspectives on Imitation: From Neuroscience to Social Science Vol.2: 55-77, Cambridge, MA: MIT.

Meltzoff A, Moore M. (1977). Imitation of Facial and Manual Gestures by Human NeonatesScience 7 (198): 75-78.

(1994). Imitation, memory, and the representation of persons. Infant Behavior and Development 17: 83-99.

Metcalfe, J. (2000). Metamemory: Theory and data. In E.Tulving & F. I. M. Craik (Eds.), The Oxford handbook of memory (pp. 197-211). New York: Oxford University Press.

Metcalfe, J., & Shimamura, A. P. eds. (1994). Metacognition: knowing about knowing. Cambridge, MA: MIT Press.

Michotte A. (1946/ English transl. 1963) The Perception of Causality, Basic Books.

Millikan R. (1984). Language, Thought, and Other Biological Categories: New Foundations for Realism, Cambridge: MIT Press.

A.D. Milner and M.A. Goodale, Separate visual pathways for perception and action, Trends in Neuroscience 15 (1) (1992), pp. 20–25.

J.P. Mitchell, M.R. Banaji and C.N. Macrae, The link between social cognition and self-referential thought in the medial prefrontal cortex, Journal of Cognitive Neuroscience 17 (8) (2005), pp. 1306–1315

J.P. Mitchell, C.N. Macrae and M.R. Banaji, Dissociable medial prefrontal contributions to judgments of similar and dissimilar others, Neuron 50 (2006), pp. 655–663

Moore C, Jarrold C, Russell J, Lumb A, Sapp F, MacCallum F (1995) Conflicting desire and the child's theory of mind. Cognitive Development 10 (4): 467-482.

Morsella, E., Bargh, J.A., & Gollwitzer, P.M. (Eds.) (2009). Oxford Handbook of Human Action. New York: Oxford University Press.

Nagel T. (1974). What is it like to be a bat? The Philosophical Review LXXXIII, 4 (October 1974): 435-50

(1986). The View From Nowhere, Oxford: Oxford University Press.

Nelson, T. O., and Narens, L.: 1992, ‘Metamemory: a theoretical framework and new findings’, in T. O. Nelson ed., Metacognition, Core Readings, 117-130.

Newton, E., (1990), Overconfidence in the Communication of Intent: Heard and Unheard Melodies, Unpublished doctoral dissertation, Stanford University (described in Goldman, 2006)

Nisbett R, Ross L. (1980). Human Inference: Strategies and Shortcomings of Social Judgment. Englewood Cliffs, NJ: Prentice Hill.

Noe A (2001) Experience and the active mind. Synthese 129: 41-60.

Oberman L, Hubbard E, McCleery J, Altschuler E, Ramachandran V, Pineda J. (2005). EEG evidence for mirrow neuron dysfunction in autism spectrum disorders. Cognitive Brain Research 24: 190-98.

Oberman L, Winkielman P, Ramachandran S. (2007) Face to face: blocking facial mimicry can selectively impair recognition of emotional expressions. Social Neuroscience 2007, 2 (3-4), 167-178.

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

O’Regan K and Noe A. (2001). What it is like to see: a sensory-motor theory of perceptual experience. Synthese 129: 79-103.

Pacherie E (2006). Towards a dynamic theory of intentions.in Does Consciousness Cause Behavior? An Investigation of the Nature of Volition (2006) 145-167 [ijn_00353954 − version 1]

(2005). Perceiving intentions.in A Explicação da Interpretação Humana, (2005) 401-414 [ijn_00353955 − version 1]

Pacherie E, Dokic J. (2006). From mirror neurons to joint actions. Cognitive Systems Research Volume 7, Issues 2-3, June 2006, Pages 101-112

Parsons, L. M. (1994). Temporal and Kinematic Properties of Motor Behavior Reflected. in Journal of Experimental Psychology: Human Perception and Performance Vol. 20, No. 4, 709-730

(1987). Imagined spatial transformation of one's body. Journal of Experimental Psychology: General, 116: 172-191.

L. M. Parsons, P. T. Fox, J. H. Downs, T. Glass, T. B. Hirsch, C. C. Martin, P. A. Jerabek and J. L. Lancaster, Use of implicit motor imagery for visual shape discrimination as revealed by PET. Nature 375 (1995), pp. 54–58

Perner, J. (1996). Simulation as Explicitation of Predication-Implicit Knowledge about the Mind: Arguments for a Simulation-Theory Mix, Carruthers P. and Smith P. 1996: 90-104

(1991). Understanding the Representational Mind. Cambridge, MA: MIT.
Perner J, Brandtl J. (2005). File change semantics for pre-schoolers: alternative naming and belief understanding. Interaction studies (Print) 6:33, 483-501.

Perner, J., Baker, S., & Hutton, D. (1994). Prelief: the conceptual origins of belief and pretence. In C. Lewis and P. Mitchell (Eds.), Children's early understanding of the mind: origins and development (pp. 261–286).

Perner J, Howe D (1995). 'He thinks he knows': And more developmental evidence against simulation (role-taking) theory. In Davies and Stone. (1995).

Perner, J. & Kühberger, A. (in press). Mental simulation: Royal road to other minds? In B. Malle & S. Hodges (ed.). Other minds: An interdisciplinary examination (166-181). New York, NY: Guilford Press.

Perner J Leekam S, Wimmer H, (2005). Three year-olds' difficulty with false belief: the case for a conceptual deficit. British Journal of Developmental Psychology 5(2): 125-37.

Perner, J. & Ruffman, T. (2005). Infants' insight into the mind: how deep?, Science, 308, 214–16.

Perner J, Leekam S and Wimmer H. (1987). Three year-olds' difficulty with false belief. British Journal of Developmental Psychology 5: 125-37.

Perner J, Lopez A., 1997, "Children's understanding of belief and disconfirming visual evidence", Cognitive Development 12, 367-380.

Perner J, Wimmer H. (1983).Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition 13 (1) 103-128.

Perrett D et al. (1989). Fameworks of analysis for the neural representation of animate objects and actions. Journal of Experimental Biology 146: 87-113.

Peterson, C., Siegel, M., 1995, "Deafness, conversation, and theory of mind," Journal of Child Psychology and Psychiatry 36, 459-474.

Petrides M, Pandya D. (1997). Comparative architectonic analysis of the human and the macaque frontal cortex. In Boller F, Grafman J (eds.), Handbook of Neuropsychology. Elsevier, Amsterdam, vol 9: 17-58.

Piaget J. (1959). Language and Thought of the Child. London: Routledge

(1957),. Etudes d'épistémologie génétique, Presses Universitaires de France: Paris.

(1950). Introduction à l'Épistémologie Génétique, Presses Universitaires de France: Paris, p.170-181

Pineda J. (2008) Sensorimotor Cortex as a critical component of an 'extended' mirror neuron system : Does it solve the development, correspondence, and control problems in mirroring? Behavioral and Brain functions 4: 47.

Pineda J, Brang D, Hecht E, Edwards L, Carey S, Bacon M, Futagaki C, Suk D, Tom J, Birnbaum C, Rork A. (2008). Positive behavioral feedback and electrophysiological changes following neurofeedback training in children with autism. Research in Autism Spectrum disorders (2008) doi:10.1016/j.rasd.2007.12.003.

Pineda J, Hecht E. (2008) Mirroring and mu rhythm involvement in social cognition: are there dissociable subcomponents of theory of mind? Biological Psychology (2008), doi:10.1016/j.biopsycho.2008.11.003.

C. A. Porro, M. P. Francescato, V. Cettolo, M. E. Diamond, P. Baraldi, C. Zuiani, M. Bazzochi and P. E. di Prampero. (1996). Primary motor and sensory cortex activation during motor performance and motor imagery: A functional magnetic resonance study. J. Neurosci. 16: 7688–7698.

Premack D, Woodruff G. (1978). Does the chimpanzee have a Theory of Mind? Behavioral and Brain Sciences 4: 515-526.

Prinz, J. (2008). Is consciousness embodied? In Ph. Robbins & M. Aydede (Eds.), Cambridge Handbook of Situated Cognition. Cambridge: Cambridge Univ. Press.

Prinz J. 2002. Furnishing the Mind: Concepts and their Perceptual Basis. Cambridge, MA: MIT Press.

Prinz, J. & Barsalou, L.W. (2000). Perceptual symbols and dynamic systems. In A. Markman & E. Dietrich (Eds.), Cognitive dynamics. Dordrecht: Kluwer Press.
Prinz, J. & Barsalou, L.W. (2002). Acquisition and productivity in perceptual symbol systems. In T. Dartnall (Ed.), Creativity, cognition and knowledge. Westport, Connecticut: Praeger.

Prinz W (2005). An ideomotor approach to imitation. In S. Hurley & N. Chater (Eds.), Mechanisms of imitation and imitation in animals (Perspectives on imitation: From neuroscience to social science, Vol. 1, pp. 141-156). Cambridge, MA: MIT Press.

(1990) A common coding approach to perception and action. In Neumann O and Prinz W (eds.).

(1990). Relationships between perception and action, Berlin: Springer.

Proust J. (2007). Metacognition and Metarepresentation: Is a self-directed theory of mind a precondition for metacognition? Synthese 159 (2), 271-295.

(2006) Rationality and metacognition in non-human animals. In Hurley and Nudds (2006).

(2003).03). Does metacognition necessarily involve metarepresentation ?
Behavior and Brain Sciences 26, 3 (2003) 352-352 [ijn_00139315 − version 1]

(2002) “Can ‘Radical’ Simulaton Theories Explain Psychological Concept Acquisition?”, in: Simulation and Knowledge of Action, J. Dokic, J. Proust (ed), Amsterdam: John Benjamins, 2002, 201-228.

Putnam H. (1975). The meaning of ‘meaning’. In Putnam H, Philosophical Papers vol. 2: Mind, Language, and Reality. Cambridge: CUP.

Quine W. (1990). Pursuit of Truth, Cambridge: MIT Press.

(1960). Word and Object. Cambridge: MIT Press.

Rakoczy H, Warneken, Tomasello M. (2007). “This way!” “No! That way!” –3-year olds know that two people can have mutually incompatible desires. Cognitive Development 22: 47-68.

N. Ramnani and R.C. Miall, A system in the human brain for predicting the actions of others, Nature Neuroscience 7 (2004), pp. 85–90.

Ramsey, F.P. (1927), "Facts and Propositions," Aristotelian Society Supplementary Volume 7, 153–170.

Ratcliffe M. (2007). Rethinking Commonsense Psychology: A critique of folk psychology, theory of mind and simulation. NY: Macmillan.

Ravenscroft, Ian, "Folk Psychology as a Theory", The Stanford Encyclopedia of Philosophy (Fall 2008 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2008/entries/folkpsych-theory/>.

(2003). Simulation, Collapse and Humean Motivation, Mind and Language 18: 162-174.

Reddy V. (1991). Playing with others’ expectations: teasing and mucking about in the first year. In: Whiten (1991).

Repacholi B, Gopnik A. (1997). Early Reasoning about Desires: Evidence from 14- to 18-month-olds. Developmental Psychology 33 (1): 12-21.

Rizzolatti G, Sinagaglia C. (2008) Mirrors in the Brain: How our minds share actions and emotions. Oxford: OUP.

Rizzolatti, G., Fadiga, L., Fogassi, L., & Gallese, V. (1996). Premotor cortex and the recognition of motor actions. Cognitive Brain Research, 3, 131-141

Rosch E (1978). Principlies of categorization. In Rosch E, Lloyd B (eds.). Cognition and Categorization. Hillsdale: Lawrence Erlbaum.

(1975). Universals and cultural specifics in human categorization. In Brislin S et al (eds.). Cross Cultural Perspectives on Learning. New York: Halsted.

Rosch E, Mervis C. (1975). Family resemblances: studies in the internal structure of categories. Cognitive Psychology 7: 573-605.

Rosch E, Mervis C, Gray W, Johnson D, Boyes-Braem P. (1976). Basic Objects in Natural Categories. Cognitive Psychology 8: 382-439.

Ross, L., Greene, D., and House, P. (1977). The false consensus effect: an egocentric bias in social perception and attribution processes. Journal of Personality and Social Psychology 13: 279-301.

Roth D, Leslie A. (1998). Solving belief problems: toward a task analysis. Cognition 66: 1-31.

M. Roth, J. Decety, M. Raybaudi, R. Massarelli, C. Delon-Martin, C. Segebarth, S. Morand, A. Gemignani, M. Décorps and M. Jeannerod. (1996). Possible involvement of primary motor cortex in mentally simulated movement. A functional magnetic resonance imaging study. NeuroReport 7: 1280–1284.

Russell J, Saltmarsh R, Hill E. (1999). What do executive factors contribute to the failure on false belief tasks by children with autism? Journal of Child Psychology and Psychiatry and Allies Disciplines 40, 859-68.

Russell, P., Hosie, J., Gray, C., Hunter, N., Banks, J., and Macauley, D., 1998, “The Developmental theory of mind in deaf children”, Journal of Child Psychology and Psychiatry 39, 905-910.

Ryle G. (1949). The Concept of Mind. London, Hutchinson.

Sabbagh M, Moses L, Shiverick S. (2006) Executive Functioning and Preschoolers’ Understanding of False Beliefs, False Photographs and False Signs. Child Development vol. 77 (4: 1034-49.

Sabbagh M, Taylor M. (2000). Neural Correlates of theory-of-mind reasoning: an event-related potential study. Psychological Science 11: 46-50.

R. Saxe, Against simulation: The argument from error, Trends in Cognitive Sciences 9 (2005), pp. 174–179.

Scheler M. (1954). The Nature of Sympathy. (Trans Barnes H) London : Routledge.

Schlick M. (1918). Allgemeine Erkenntnistheorie, Berlin: Springer.

Scholl and Termoulet (2000) Perceptual Causality and Animacy. In: Trends in cognitive sciences Vol. 4, Issue 8, 299-309

Searle J. (1983). Intentionality: an essay in the philosophy of mind. Cambridge University Press.

Segal G. (1996). The modularity of theory of mind. IN Carruthers and Smith (1996).

Sellars W. (1956). Empiricism and the philosophy of mind. In: Feigl, H. and Scriven, M. (eds) The Foundations of Science and the Concepts of Psychology and Psychoanalysis: Minnesota Studies in the Philosophy of Science, Vol. 1. Minneapolis: University of Minnesota Press: 253-329.

Shastri, L. & Ajjanagadde, V. (1993). From simple associations to systematic reasoning: A connectionist representation of rules, variables and dynamic bindings using temporal synchrony. Behavioral and Brain Sciences, 16, 417-494.

Shimamura A. P. (2000). Toward a cognitive neuroscience of metacognition. Consciousness and Cognition, 9, 313-323.

Slaughter V. (1998). Children's understanding of pictorial and mental representations. Child Development 69: 321-32.

Smith E, Medin D, Rips L. (1984). A psychological approach to concepts: comments on Georges Rey's 'Concepts and Stereotypes.' Cognition 17: 265-274.

Smith J. D. (in press). Animal metacognition and consciousness. In Cleeremans A., Bayne T.,

Smith, J. D., Shields, W. E., & Washburn, D. A. (2003). The comparative psychology of uncertainty monitoring and metacognition. The Behavioral and Brain Sciences, 26 317-371.

Smith, J. D., Schull, J., Strote, J., McGee, K., Egnor, R., & Erb, L. (1995). The uncertain response in the bottlenosed dolphin (Tursiops truncatus). Journal of Experimental Psychology: General, 124, 391-408

Smith, L.B. & Jones, S. (1993). Cognition without concepts. Cognitive Development, 8, 181-188.

Sommerville J, Woodward A, and Needham A. (2005). Action experience alters 3-month-old infants' perception of others' actions, Cognition 96: B1-B11.

Sperry R. (1950). Neural basis of the spontaneous optokinetic response produced by visual inversion, Journal of Comparative and Physiological Psychology 43: 482–489.

Spivey, M. et al. (2000) Eye movements during comprehension of spoken
scene descriptions Proceedings of the 22nd Annual Conference of the
Cognitive Science Society, Erlbaum, pp. 487–492

Sprengelmeyer, R., Young, A. W., Schroeder, U., Grossenbacher, P. G., Federlein, J., Buttner, T., and Przuntek, H., 1999, "Knowing no fear", Proceedings of the Royal Society, series B: biology 266: 2451-2456.

Stanfield R.A. & Zwaan R.A. (2001). The effect of implied orientation derived From verbal context on picture recognition. Psychological Science, 12,153-156.

Stich S, Nichols S.(2003). How to read your own mind: a cognitive theory of self-consciousness. In Consciousness: New Philosophical Essays, eds. Q. Smith and A. Jokic. Oxford University Press, 157-200.

(1995). Second thoughts on simulation. In Davies and Stone (1995a).

(1992). Folk Psychology: Simulation or Theory. Mind & Language, 7, no. 1, 35-71.
Stone T, Davies M (1996) The mental simulation debate: a progress report. In Carruthers and Smith (1996)

Stich S, Ravenscroft I. (1994). What is Folk Psychology?. Cognition 50: 447-68.

Stueber K. (2006). Rediscovering Empathy: Agency, folk psychology, and the human sciences. Bradford, MIT Press: Cambridge MA and London.

Tager-Flusberg H, Sullivan K. (2000). A componential view of theory of mind: evidence from Williams Syndrome. Cognition 76:59-89.

Tollefsen D. (2005). Let’s pretend!: Children and Joint Action. Philosophy of the Social Sciences 35: 75-97.

Tomasello M. (2008). Origins of Human Communication, Cambridge MA: MIT Press.

(1999). The Cultural Origins of Human Cognition, Harvard University Press: Cambridge

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 28:5:675-691

Tomasello M, Rakoczy H, (2003). What makes human cognition unique? From individual to shared to collective intentionality. In Mind and Language 18 (2), April 2003.

Trevarthen C. (1979). Communication and cooperation in early infancy: a description of primary intersubjectivity. In Bullowa M (ed) Before Speech. Cambridge: Cambridge Univ. Press.

Umiltà M, Kohler E, Gallese V, Fogassi L, Fadiga L, Keysers C, Rizzolatti G. (2001). I know what you are doing: a neurophysiological study. In Neuron 32: 91-101.

Van Boven, L., Loewenstein, G., 2003, „Social projection of transient drive states,“ Personality and Social Psychology Bulletin 29(9): 1159-1168.arela, F.J. & Thompson, E. & Rosch, E. (1992). The embodied mind – Cognitive science and human experience. Cambridge, Mass.: The MIT Press.

De Vignemont F and Fournet P. (2004). The sense of agency: a philosophical and empirical review of the ‘who’ system. Consciousness and cognition 13, 1-19.

K. Vogeley, P. Bussfield, A. Newen, S. Herrmann, F. Happe and P. Falkai et al., Mind reading: Neural mechanisms of theory of mind and self-perspective, Neuroimage 14 (2001), pp. 170–181

K. Vogeley, M. May, A. Ritzl, P. Falkai, K. Zilles and G.R. Fink, Neural correlates of first-person perspective as one constituent of human self-consciousness, Journal of Cognitive Neuroscience 16 (2004), pp. 817–827

Wellman H. (1990). The Child's Theory of Mind. Cambridge MA. MIT Press.

(1991). From beliefs to desires: acquisition of atheory of mind. In Whiten (ed.) (1991).

Wellman H. and Bartsch, K. (1988), Young Children's reasoning about beliefs. Cognition, 30, 239-77

Wellman H and Woolley J. (1990). From simple desires to ordinary beliefs: The early development of everyday psychology. Cognition 35,: 245-75.

Whiten, A. (ed.) (1991), Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading, Oxford: Basis Blackwell.

Wilken P. (Eds.), The Oxford companion to consciousness. Oxford: Oxford University Press.

Winsberg E. (2003). "Simulated Experiments: Methodology for a virtual world", Philosophy of Science, 70, p. 105-125.

Wisniewski, E. (1997). When concepts combine. Psychonomic Bulletin and Review 4: 167-183.

Wittgenstein L (1953). Philosophical Investigations. Anscombe E trans. Ny: Macmillan.

Wolpert, D. M., Doya, K., & Kawato, M. (2003). A unifying computational framework for motor control and social interaction. Philosophical Transactions of the Royal Society, LondonB, 358, 593-602

Wynn, K. (1992) "Addition and Subtraction by Human Infants." Nature 358:749-750.

Wundt W. (1896). Grundriss der Psychologie, Leipzig: Engelmann.

B. Wuyam, S. H. Moosavi, J. Decety, L. Adams, R. W. Lansing and A. Guz. (1995). Imagination of dynamic exercise produced ventilatory responses which were more apparent in competitive sportsmen. J. Physiol. 482: 713–724.

Zahavi D. (2008). Simulation, projection and empathy. Consciousness and Cognition 17: 514-522.

Zaitchik D. (1990). When representations conflict with reality: The pre-schooler's problem with false beliefs and 'false' photographs. Cognition 35: 41-68.

Zhong C., Leonardelli, G. (2008). Cold and Lonely: Does Social Exclusion Literally Feel Cold? Psychological Science 19 (9), 838-842.

Zhong C., Liljenquist K. (2006) Washing away your sins: threatened morality and physical cleansing. Science 8 September 2006: Vol. 313. no. 5792, pp. 1451 - 1452

Curriculum Vitae

Education

PhD Project "What is folk psychology and who cares?" in progress since 2006, Department of Philosophy. University of Vienna, within the doctoral program (*Initiativkolleg*) "Natural Sciences in Historical Context".

M.A. September 2006, University of Tübingen, in Philosophy (Major Subject/ *Hauptfach*), grade: very good (*sehr gut*); Rhetoric (Minor Subject/ *Nebenfach*), grade: very good (*sehr gut*); Cultural Studies (=*empirische Kulturwissenschaft*) (Minor Subject/ *Nebenfach*), grade: very good (*sehr gut*).

B.A. June 2000, Wesleyan University (Middletown, Connecticut, USA) in Philosophy

Academic Work Experience

2007-2008: Spokesman (Sprecher) for the doctoral students within the doctoral program (*Initiativkolleg*) "Natural Sciences in Historical Context".

2006-Present: Graduate Assistant (*Kollegassistent*) in the doctoral program (*Initiativkolleg*) "Natural Sciences in Historical Context" at the University of Vienna.

2002-2006: Student Assistant (*Hilfskraft*) to Professor Dr. Michael Heidelberger, Chair for Logic and Scientific Theory in the Philosophy Department of the University of Tübingen.

2003- 2006: Translator for Prof. Dr. Manfred Frank, Philosophy Department of the University of Tübingen.

2003-Present: Free-lance translator of philosophical articles

Teaching Experience

Summer 2009: Tutor, Introductory Seminar: History of the Vienna Circle

Winter 2008-09: Instructor, Introductory Seminar: Introduction to Philosophy of Science

Summer 2007: Tutor, Guided Reading Course: Introduction to the History of Science

February-March 2008: Organizer of Interdisciplinary Reading Group ("Philosophy of Mind and Developmental Psychology") at MPI for Evolutionary Anthropology in Leipzig

Summer 2006: Tutor, Proseminar: Introduction to Logic

Summer 2003: Tutor, Proseminar: Introduction to Philosophy of Science

2000-2001: English Teacher, Berlitz Sprachschule, Stuttgart.