Simulation Theory versus Theory Theory

Theories concerning the Ability to Read Minds

Diplomarbeit
zur Erlangung des akademischen Grades eines Magisters
an der Geisteswissenschaftlichen Fakultät der
Leopold-Franzens-Universität Innsbruck

eingereicht bei Herrn Univ.-Doz. Dr. Hans Kraml

von Martin Michlmayr <tbm@cyrius.com>

Innsbruck, März 2002

Contents

1 Introduction

The philosophy of mind debate is one of the most important topics in contemporary philosophy. The problems are so extensive and complex that the debate has been going on for a very long time and yet there is no end in sight. Many problems have been solved over the years but at the same time new questions have been posed. Furthermore, the focus of the debate itself has shifted in the last few decades.

Greenwood (1991) claims that in the past "most philosophers were almost exclusively concerned with the question of whether the qualitative aspects of sensations such as pain or sense impressions could be reduced to brain states" (p. 1). In recent years, many philosophers have lost interest in the problem of *qualia* because neuroscience has made such tremendous progress in this area (Churchland, 1990). Philosophers are increasingly involved in a debate about social interaction and intentional psychological states such as beliefs, desires, emotions, and motives.

The importance of social interaction has been recognized in various disciplines in recent years. In fact, it has even been suggested that "the evolution of intelligence in primates that ultimately led to human beings was driven in part by the demands of social information processing" (Barresi & Moore, 1996, p. 107). Humphrey (1984), who proposed that intelligence evolved in order to support organisms living in complex groups, coined the metaphor of *social chess* to illustrate his ideas. He claims that:

> Like chess, social interaction is typically a transaction between social partners. One animal may, for instance, wish by his own behavior to change the behavior of another; but since the social animal is himself reactive and intelligent the interaction soon becomes a two-way argument where each "player" must be ready to change his tactics – and maybe his goals – as the game proceeds. Thus, over and above the cognitive skills which are required merely to perceive the current state of play (and they may be considerable), the social gamesman, like the chess-player, must be capable of a special sort of forward planning. Given that each move in the game may call forth several alternative responses from the other player this forward planning will take the form of a decision tree, having its root in the current situation and branches corresponding to the moves considered in looking ahead at different possibilities. It asks for a level of intelligence which is, I submit, unparalleled in any other sphere of living. (p. 20–21)

Humphrey's metaphor of social chess is indeed a good illustration of the theory that in order to deal with increasing social demands, primates, especially humans, have evolved a system that is used for understanding, predicting and manipulating the behavior of others. Using a phrase from Premack & Woodruff (1978), who worked with chimpanzees, this system is usually called a *theory of mind*. Simon Baron-Cohen, who is famous for his work on autism, calls the ability *mindreading*.

1.1 Functions of Theory of Mind

We take our theory of mind (ToM) for granted and usually do not think about the influence it has on our lives. Just like breathing, the functions of ToM happen automatically without requiring volitional actions. When we see someone walking to his car and then rummaging in his pockets, we

assume that he is *searching for his keys.* This interpretation would not be possible without intentional understanding which is made possible by our theory of mind. Baron-Cohen (1995) underlines the importance of a fast and effective mindreading system from an evolutionary perspective:

> [I]magine that you are an early hominid, and that another early hominid offers to groom you and your mate. You need to reason quickly about whether you should let him approach. [...] Making inferences about whether his motives are purely altruistic or whether he might be deceitful is a reasoning strategy that you can apply in time to react to a social threat. (p. 25)

This example shows that from an evolutionary point of view theory of mind can be seen as a successful surviving strategy. The fast-acting mindreading system gives you immediate information about the intentions of the people around you, which in turn allows you to act quickly and safely. Furthermore, as observed previously in the description of the chess metaphor, theory of mind is a crucial prerequisite for the evolution and development of complex social systems and interactions. In particular, the mindreading system has these functions:

- Comprehend and explain: theory of mind allows us to see a meaning in the behavior and actions of other people. Without it, we would be confused by other people's actions and overwhelmed by the complexity of daily life. Theory of mind creates an order in life by giving everything a purpose and meaning.

- Predict: theory of mind also gives us the ability to predict other people's behavior. This is a requirement for dealing with other people, or, as Churchland (1991) puts it, "if one cannot predict or anticipate the behavior of one's fellows at all, then one can engage

in no useful commerce with them whatever" (p. 57). Furthermore, by making the world predictable, much complexity is taken. When one knows that something is going to happen as well as the reason why it will happen one can adjust to the situation in advance.

- Manipulate: one can use theory of mind to influence and manipulate the behavior of others by controlling the information available to them. However, this can only be done when one perceives the other's goals, desires and beliefs through one's mindreading ability.

We have now learned what role theory of mind plays. However, since we all posses theory of mind and take it for granted, it is difficult for us to see when we actually use this facility. Fortunately, we can increase our understanding using a method common in cognitive science and neuroscience – we can move out of our usual perspective and take a look at the situation when theory of mind is not functioning properly. Autism will act as an example for this case.

1.2 Living without Mindreading

Autism is a developmental disorder of the brain which exists from birth and lasts the whole life. Although most research on autism has only been conducted during the last few decades, Uta Frith (1989) speculates that the disorder is not new. Frith has found evidence for autism throughout history and cites the "Blessed Fools" of Old Russia as an example. They were described as showing "bizarre behaviour, innocence, and lack of social awareness" (Happé, 1994, p. 7).

Autism has first been described independently by Kanner (1943) and Asperger (1944). They highlighted features such as strange social interaction, desire for the preservation of sameness, excellent rote memory,

delayed echolalia and oversensitivity to stimuli. Although autism is a biologically based disorder with a strong genetic component, the diagnosis is still based on behavioral criteria (Happé, 1999). The criteria are usually based on Wing's triad, which consists of (Happé, 1994, p. 20):

- qualitative impairment in reciprocal social interaction
- qualitative impairment in verbal and nonverbal communication and in imaginative activity
- markedly restricted repertoire of activities and interests

In his book, Baron-Cohen (1995) suggests that "children and adults with the biological condition of autism suffer, to varying degrees, from mindblindness" (p. 5). In other words, their mindreading ability is impaired. He starts his book by asking his readers to "imagine what your world would be like if you were aware of physical things but blind to the existence of mental things" (p. 1). This is a very difficult task since we attribute beliefs and desires to each other unselfconsciously all the time. Baron-Cohen continues by giving an example of a simple human act, namely:

> John walked into the bedroom, walked around, and walked out.

Baron-Cohen then asks how we would make sense of this sentence. He offers some typical explanations (p. 1) from humans with the mindreading ability, the so called "mindreaders":

- Maybe John was **looking** for something he **wanted** to find, and he **thought** it was in the bedroom.
- Maybe John **heard** something in the bedroom, and **wanted to know** what had made the noise.

- Maybe John **forgot** where he was going: maybe he really **intended** to go downstairs.

We as mindreaders typically use mental-state words (printed above in boldface) to make sense of John's act. However, people with mindblindness appear to have only limited access to such explanations. They can merely give simple statements about temporal regularities ("he does it everyday"). If this explanation fails – which it most likely will – not many simple, readily available, plausible, non-mentalistic explanations are left.

It is therefore not suprising that Frith (1996), who describes the inability to attribute mental states as "tantamount to not differentiating between the world of objects (with physical states) and the world of persons (with mental states)" (p. 65), claims that the social world must be very alienating for people with limited mindreading ability. Mindreaders are sometimes lost when they enter cultures in which acts and gestures have different meanings. People with mindblindness always live in a social world without meaning.

1.3 Theories

The cases of people with some form of mindblindness, and theoretical and empirical studies about the evolution of primates underline the importance of the mindreading device. There is no dispute that our complex social system and interaction would not be possible without an effective theory of mind. However, there are various theories offering different explanations how exactly the mindreading mechanisms work. These theories are usually grouped into two categories, *theory theory* (TT) and *simulation theory* (ST).

Theory theory is the theory which has been predominant in the last few decades. It states that our understanding of the mind is based on a folk psychological theory which consists of a framework of concepts which is "roughly adequate to the demands of everyday life" (Churchland, 1991, p. 51). Explanations are derived from a set of laws and rules that "connect the explanatory conditions with the behavior explained" (Churchland, 1990, p. 207). There is no consensus among theory theorists of which type theses laws and rules are. However, generally speaking, early theories have proposed a set of explicit laws comparable to those of a full-fledged science. Probably due to the influence of Thomas Kuhn, who pointed out that implicit assumptions are part of science, it has recently become more common to propose implicit or tacit laws.

Some philosophers and scientists believe that this folk psychological theory is learned as the child grows up (e.g. Churchland, 1991), while others hold a nativistic attitude. Carruthers (1996), for example, supports the nativistic theory because he finds it "puzzling how [learning the theory] can take place without any explicit teaching or training" (p. 22).

In recent years, increasing attention has been given to the alternative account of mindreading known as simulation theory (ST). ST denies that we come to an understanding of others through the deployment of a theory. Instead, it suggests that we use "the resources of [our] own minds to simulate [...] others." (Davies & Stone, 1995a, p. 3). ST has roots in the *Verstehen* methodology of Dilthey (Heal, 1995, p. 39) and in *empathizing*. By putting yourself in the shoes of someone else, you can simulate them and come to predictions and explanations.

It is important to note that in the debate of theory theory and simulation theory terms often have different meanings depending on who uses them. Sometimes, the term "folk psychology" (FP) is used to refer to the

mindreading ability per se, regardless of whether the ability is accounted for using theory theory or simulation theory. In other cases, the term "folk psychology" is only used in the context of theory theory. It is not used for simulation theory since the use of the word "psychology" suggests that a theory lies behind the mindreading ability and this is exactly what simulation theory denies. In this paper, I will adhere to the latter usage and use "folk psychology" strictly for theory theory. When the ability to understand and predict others is meant, I will use the terms "mindreading" or "theory of mind".

In the following, I am going to describe both of these theories in more detail in order to investigate what each has to offer in terms of explanation of the mindreading ability. The predominant theory theory will be evaluated first to see which solutions and explanations have been proposed in the last few decades. Afterwards, I am going to discuss simulation theory in more detail. Different ST theories will be presented and contrasted. Furthermore, some attention will be given to current research in cognitive science and neuroscience which are connected to ST. In the end, insights won by studying theory of mind will be summarized and conclusions will be drawn.

2 Theory Theory

Theory theory claims that we posses a folk psychological ability which rests upon knowledge of a theory. Just like other folk theories, such as folk physics, it enables us to master our daily lives successfully. Although we use this theory constantly throughout the whole day, we are not actually aware of the laws of which the theory is composed. The theory therefore is of implicit and tacit nature.[1]

Blackburn (1995) tries to give a description of the relation between tacit knowledge and theories. He states:

> If we are good at something [...] then we can be thought of as making tacit (very tacit) use of some set of principles that could, in principle, provide a description of a device, or possibly a recipe for the construction of a device, that is also good at it. (p. 275)

This account, which resembles the ideas of the mathematician and computer pioneer Alan Turing, outlines a very weak conception of a theory. If it was sufficient to find a theoretical description, there would be no dispute between TT and ST at all. After all, it is plausible to develop a theoretical representation of simulation and empathizing. Although Blackburn's thoughts do not give an adequate description of theories, they raise the question what a theory really is.

[1] There are theories which propose an explicit folk psychological theory, too. However, implicit and tacit theories appear to be more common these days; therefore, explicit theories are not covered here.

2.1 Is TT a Theory?

The theoretical view of ToM holds the position that our common-sense terms for mental states are part of a theoretical framework, namely folk psychology (FP), which is embedded in our common-sense understanding (Churchland, 1992). The theory can be seen as containing a "large number of universally quantified conditional statements, conditions with the conjunction of the relevant explanatory factors as the antecedent and the relevant explanandum as the consequent" (Churchland, 1991, p. 52–53). These "laws" therefore express the relations between the various properties and entities postulated by the theory. Churchland (1990) states that:

> Each of us understands others, as well as we do, because we share a tacit command of an integrated body of lore concerning the lawlike relations holding among external circumstances, internal states, and overt behavior. Given its nature and functions, this body of lore may quite aptly be called "folk psychology." (p. 207)

Churchland (1990, 1991) is one of the most forceful supporters of the idea that folk psychology is an "empirical theory that is subject to the same canons of empirical evaluation as any other" (Davies & Stone, 1995b, p. 7). According to Churchland (1991) the theory may be evaluated for its virtues and may be rejected fully if it fails the measure of evaluation. Indeed, the evaluation is very simple. The framework of folk psychology as an empirical theory is successful if it helps explaining and predicting human behavior at large. Churchland (1991) criticizes those who ask for a justification of folk psychology and suggests that "folk psychology is justified by what standardly justifies *any* conceptual framework: namely, its explanatory, predictive, and manipulative success" (p. 61).

One common objection to the framework of folk psychology is that it does not have the character of genuine causal/explanatory laws. Rather, the theory and its laws have some other, less empirical status, such as that of normative principles. Churchland (1991) on the other hand defends the empirical and theoretical nature of folk psychology. He groups folk psychological concepts into two broad classes. First, there are fully intentional concepts expressing various propositional attitudes, such as beliefs and desires. Second, there are non-intentional or quasi-intentional concepts expressing all other mental states, such as grief, fear, pain and hunger. Churchland (1991, p. 53) lists some typical generalizations for the latter type, including:

- A person who suffers severe bodily damage will feel pain.
- A person who suffers a sudden sharp pain will wince.
- A person denied food for any length will feel hunger.

He furthermore observes that these generalizations, "and thousands more like them" (p. 53), are causal/explanatory in character. They are used for simple explanations and have a great importance in folk psychology. Churchland claims that "concepts of this simple sort carry perhaps the major part of the folk psychological burden" (p. 53). Furthermore, he suggests that concepts expressing propositional attitudes are empirical since "on the basis of presumed information about the current cognitive states of the relevant individuals, one can nonaccidentally predict at least some of their future behavior some of the time. But any principle that allows us to do this – that is, to predict one empirical state or event on the basis of another, logically distinct, empirical state or event – *has* to be empirical in character" (p. 54). These logical relations are akin to those in high-grade theoretical frameworks in science. Summing up, the

simpler parts of folk psychology are causal and the more complex parts posses the same sophisticated logical structure as powerful theories in science.

One implication of the theoretical nature of folk psychology is that it is not dependent on knowledge of one's own mind, and, more broadly, that it is not dependent on human psychology. In other words, TT does not require a specific psychology from the attributor. Anyone, including Martians (Churchland, 1990), who knows the laws and propositions which make up folk psychology can form predictions and explanations through proper reasoning and use of the laws. Goldman (unpublished) describes this feature of TT in more detail:

> [T]he body of knowledge used to make attributions would be just as effective, or accurate, in the hands of an alien creature as it [is] in the hands of a human attributor. If the alien is just as competent at wielding this body of knowledge as a human, the fact that the alien might have a radically different psychology from his human target would make no difference. (p. 4)

This characteristic can be seen as a great advantage of TT. For one, it would imply that artificial intelligence (AI) should be possible in principle and that we can expect computers and robots which understand folk psychology.[2] Furthermore, a merely theoretical framework is often much easier to study and investigate since it can largely be explored without having to pay attention to the attributor's specific features and characteristics.

[2] In fact, there are projects, such as Cyc (http://www.cyc.com/), which try to teach computers folk physics and folk psychology.

2.2 TT and 'Reality'

Folk psychology (FP) is supposed to be a very old theory and like other folk theories it has changed little or not at all since ancient times (Carruthers, 1996; Churchland, 1991).[3] Since FP is an empirical theory, it is possible in principle that it is false. This means that it could be that "its principles are radically false and that its ontology is an illusion" (Churchland, 1990, p. 210). Any part of FP might be "overthrown and replaced by some other doctrine" (Dennett, 1991, p. 135). In fact, insights won through the huge progress in recent times in neuroscience suggest that FP most likely does not describe cognitive processes adequately.

However, it is important not to draw wrong conclusions from these observations. It does not necessarily mean that FP will be given up for a more correct, scientific theory. Dennett (1991), for example, underlines that FP is here to stay:

> What I want to stress is that for all its blemishes, warts, and perplexities, folk psychology is an extraordinarily powerful source of prediction. It is not just prodigiously powerful but also remarkably easy for human beings to use. (p. 135)

What one has to keep in mind is that the debate whether theory of mind is based on a theory or on simulation is the question of how we perform mentalizing in our everyday life. The question whether this layman understanding describes mental concepts properly in a scientific sense is on a totally different level. There is only one position which suggests that folk psychological understanding and a scientific theory cannot exist at the same time, namely eliminativism.

There are various theories about the relation between people's understanding of qualitative states and their mentalizing ability, and the

[3] Our scientific and philosophical understanding of folk psychology, however, has changed dramatically and is still adapting to new insights.

mental states in the brain as discovered and described by neuroscience on the other side. The identity-theorist thinks that it will be possible to reduce FP to neuroscience smoothly. The ontology of FP will be preserved and relations to scientific entities be shown. The dualist on the other hand thinks that such a reduction will not be possible since non-physical entities are involved. The eliminative materialists think that FP will be replaced by a better theory because they see FP as a "radically inadequate account of our internal activities, too confused and too defective to win survival through intertheoretic reduction" (Churchland, 1990, p. 209–210).

Initially I have claimed that theory theory was an empirical theory which is not immune to revision. However, revision does not necessarily mean that the whole theory will be replaced by a better one, as it is the goal of the eliminative materialist. FP could be improved within its framework. Although FP might not be accurate on the level of brain states, it *does* describe the empirical circumstances in the user's environment properly and therefore allows adequate explanations and predictions. After all, it is not the goal of FP to give a proper account of neuroscience, but to allow mindreading.

2.3 Acquiring Folk Psychology

The question how this mindreading facility is actually acquired remains, however. As seen previously, folk psychology has much in common with professional science. There are differences as well, though. Folk psychology is "not learned by way of explicit formal teaching; nor is it written up in text book form" (Davies & Stone, 1995b, p. 12). Churchland suggests that the principles making up the theoretical framework of FP are

learned "at mother's knee, as we learn our language" (cited in Davies & Stone, 1995b, p. 10). In other words, the laws are acquired as we grow up by implicit "teaching" through others (especially our mothers) and by learning step by step by living together with other people and having social interactions all the time.

Carruthers (1996) opposes this idea and suggests instead that the folk psychological theory is given innately, rather than acquired through learning of any sort. He has two main arguments to support his nativistic thesis. First, if one pictures young children as little scientists constructing a theory, it is strange that they all come to the same theory at the same time (which they in fact do at about the age of four). Second, if FP on the other hand is learned from adults, how should that take place without explicit teaching? Furthermore, Carruthers observes that the folk psychological theory has remained invariant across cultures and historical eras, which he finds strange if the theory is a cultural construct. Although Carruthers does not mention it, one might also view autism as supporting the nativistic theory. However, this disorder can just as well be accounted for with a non-nativistic explanation. It could be that the facility which normally allows children to learn the theory is impaired.

The question whether FP is learned or given at birth can certainly not be answered satisfactorily at this point. However, in my opinion, there are some arguments against Carruthers' thesis. He himself claims that the suggestion that FP might be innate "is not at all implausible, given the crucial role that it plays in facilitating communication and social co-operation in highly social creatures such as ourselves" (Carruthers, 1996, p. 23). This argument can, however, be directed against Carruthers as well since – with the same plausibility as giving FP innately – nature could have created a very effective device to allow the *learning* of such a theory. This in turn could explain why children come to the

same theory at about the same time. The advantage of this approach over hardcoding a theory is that it is more flexible and allows the theory to be adapted more easily. The question how FP can be learned from adults without formal teaching is similar to the one asked in linguistics. The fact that people of different nationalities speak different languages obviously suggests that language itself is not innate. Why then should a folk psychological framework be given innately?

Interestingly, the study of FP shows further parallels to problems in linguistics. Although we are fluent in our mother tongue and can solve many linguistic problems without thinking, we fail completely when we try to describe the laws which determine our language. The same goes for FP, which Churchland (1991) describes very eloquently:

> If one's capacity for understanding and predicting the behavior of others derives from one's internal storage of thousands of laws or nomic generalizations, how is it that one is so poor at enunciating the laws on which one's explanatory and predictive prowess depends? It seems to take a trained philosopher to reconstruct them! (p. 61–62)

This is not necessarily an argument for us not really using laws and sentences at all (and that therefore simulation theory should be used to explain ToM). It merely shows the tacit nature of the folk psychological framework which I discussed at the beginning of this chapter.

2.4 Testing for Theory of Mind

While the question how exactly theory of mind is acquired remains open, it appears that children at about the age of four show a folk psychological understanding which is, at least in the main areas, comparable to that

of adults. One could ask for a test which shows how far theory of mind is developed. In fact, since the work with chimpanzees by Premack & Woodruff (1978) "false belief has been considered a kind of litmus test for the presence of a theory of mind" (Barresi & Moore, 1996, p. 118).

Wimmer & Perner (1983) developed the first false belief task for children, and carried out a widely acclaimed experiment. In the task children are presented with the following story: Maxi places some chocolate in a cupboard in the kitchen and leaves the room. While Maxi is away, another character takes the chocolate from the first cupboard and puts it in the second cupboard, and then leaves. When Maxi returns, the child is asked to predict where Maxi will look for the chocolate.[4] The correct answer of course is that Maxi will look for the chocolate were he left it (i.e. in the first cupboard).

In the experiment, two groups of children have been studied. The children in the first group had an average age of about three, the second group of roughly five. For the experiment to work properly, the experimentator of course has to ensure that the child understands the story and remembers where the chocolate has been placed. After the experiment had been performed, it has been shown that the younger group failed to answer the question correctly while children in the older group had little difficulty giving the correct answer. The older children understand that Maxi does not know that the chocolate has been moved while he was out of the room and that he will therefore act upon a false belief.

The assumption why older children get it right while younger ones do not is that

> the older children have an understanding of folk psychology that is in key respects identical to the understanding that the

[4] A similar version is that Sally puts a marble in a basket. While she is away, Anne moves the marble to a box. Again, the child has to guess where Sally will expect to find the marble.

> mature adult has, whereas the younger children lack some aspects of adult understanding. Given this assumption, the experiment becomes a diagnostic of whether a child has attained the mature state with respect to key components of the conceptual repertoire that comprises our folk psychology. (Davies & Stone, 1995b, p. 3)

Barresi & Moore (1996) try to give an account of why the task is so hard to accomplish for children under the age of four. They argue that the younger children typically fail because they answer "on the basis of the real location of the hidden object" (p. 119). Clearly, there is no "real location" independent of any agent or observer. What Barresi & Moore mean by "real location" is the content of the child's own current belief. However, it does not necessarily correspond to the real location. After all, it is imaginable that the object in the false belief task is moved again without the knowledge of the child participating in the experiment.

Barresi & Moore identify two critical aspects about the task. First, the child has to "generate a representation of the agent's epistemic relation to the situation, for which that agent's third person information is not perceptually given" (p. 119). Second, it has to image the first person information of the agent, while at the same time having a different first person intentional relation. This is critical because it implies that "neither current third person information nor current first person information are available to generate the requisite representation of the intentional relation" (p. 119). However, according to their theory, younger children can only imagine *one* component of an intentional relation. When younger children are presented with third person information, they are able to imagine the first person information. However, when third person information lacks as well, they can only generate a presentation of the other's intentional relation by using the current first person information

of *themselves*. Unfortunately, this leads to a wrong prediction. Only at about the age of four children get the "capacity to perform what amounts to a double imaginative act, whereby the child can generate, through the use of the intentional schema, a representation of an intentional relation for which both first and third person information are imagined" (p. 119).

A different account of the failure in the false belief task is Josef Perner's theory (1991) about meta-representation. He argues that one has to be capable of meta-representational thought in order to solve the false belief task. In other words, one needs to have the ability to represent someone else's act of representing the world. Perner suggests that children at the age of two are situation theorists. At about four a shift happens and children become representation theorists. This leads to the understanding that beliefs are attitudes towards representations of reality rather than towards reality itself.

The simulation theorists have a radical different explanation for the failure of children younger than four in the false belief task, as we will shortly see.

3 Simulation Theory

The history of simulation theory reaches back quite far. Simulation (or empathy) has roots in Dilthey's *Verstehen* methodology and Goldman (unpublished) argues that the great philosophers Hume and especially Kant had strong simulationist learnings. Similarly, Perner & Howes (1992) describe that simulation is an old idea in developmental psychology circles which has great importance in Piaget's psychology. In particular, simulation – known as *role-taking* or *perspective-taking* in Piaget's theory – helps young children overcome their egocentric views.

According to Fuller (1995), simulation and empathy was "killed and buried" by the positivists (p. 19). They distinguished between the context of discovery and the context of justification and claimed that empathy only belonged to the former context. While simulation can be used as a great heuristic device to suggest predictive and explanatory hypotheses, it cannot be used to justify these hypotheses – formulation and testing of generalizations have to be done for a proper justification.

However, empathy and simulation have been resurrected in the last few decades. Putnam (cited in Fuller, 1995, p. 19), for example, argues that empathy plays a role in justification of hypotheses because it "gives plausibility".

Simulation theory (ST) today has a strong influence on the philosophy of mind debate. ST suggests that we do not understand others through the use of a folk psychological theory. Rather, we use our own

mental apparatus to form predictions and explanations of someone by putting ourselves in the shoes of another person and simulating them. ST is often described as *off-line* simulation, although there are philosophers who maintain that off-line simulation is only an ancillary hypothesis of ST (see Davies & Stone, 1995a, p. 4). In off-line simulation, one takes one's own decision-making system off-line and supplies it with pretend inputs of beliefs and desires of the person one wishes to simulate in order to predict their behavior. One then lets one's decision-making system do the work and come to a prediction.

There are many variants of ST, some differing more than others. While some philosophers suggest a hybrid theory of TT and ST, others argue that ST should replace the predominant TT. Gordon, for example, who holds some of the strongest claims, suggests that simulation is fundamental to the mastery of psychological concepts and that it has ramifications for the ontology of psychological states (Fuller, 1995). While there are many varieties and different views of ST, all have in common that simulation acts as a very effective device for forming predictions and explanations. This leads to an important implication of ST. Since simulation depends on one's own mental apparatus, it is clear that ST (in contrast to TT) is attributor dependent.

3.1 ST as a 'Hot Theory'

Gordon (1996) describes theory theory as a *cold theory*. A cold methodology mainly uses intellectual processes, makes inferences from one set of beliefs to another and "makes no essential use of our own capacities for emotion, motivation, and practical reasoning" (p. 11). On the other hand, there is the *hot* methodology. It makes use of one's own moti-

vational and emotional resources and one's own capacity for practical reasoning. Simulation theory is a typical hot methodology.

In a similar way as Gordon, Goldman (unpublished) distinguishes between attributor neutral (AN) and attributor dependent (AD) heuristics. In the discussion about the characteristics of TT and the possibility of AI (see page 15), it has clearly been shown that TT is an AN heuristic. Goldman argues for AD heuristics from an evolutionary perspective. He first cites David Marr saying that "an algorithm is likely to be understood more readily by understanding the nature of the problem being solved". Then, he suggests that AD heuristics are *ecologically rational* because other human beings have "major psychological similarities to any prospective attributor" (p. 3). It is therefore very effective to use one's own mind since one can achieve accurate predictions and explanations with few cognitive resources.

Since AD heuristics are so effective, Goldman suspects that "evolution might have hit upon this kind of heuristic" (p. 3). At the same time he suggests that evolution gave us more than only one mentalizing strategy. In neuroscience it is common knowledge that one can often learn a task again after an impairment occurred using non-standard resources. Goldman names autism as a case of this for the mindreading ability. While autistic children are impaired in the mentalizing domain,[1] they can learn to use theorizing resources to perform mindreading.

While Goldman is a proponent of the AD model, he points out an important problem this approach has to face. When we put ourselves in the shoes of someone else, this simulation "does not involve the very same states in the attributor as those undergone by the target" (p. 11). Rather than having beliefs and desires, the simulator has *pretend* beliefs and *pretend* desires. The important question therefore is whether these

[1] This means, according to Goldman, that their simulation capability is impaired.

pretend states are "sufficiently similar – in psychological and perhaps neurological terms – to their genuine counterparts" (p. 11).

Goldman clearly thinks that they are sufficiently similar and presents evidence from various domains. Specifically, he names visual imagery (pretense-generated vision) and motor imagery (pretense-generated motor instructions) as supporting the similarity of real and pretense states. The experiment of Shepard & Metzler (1971) about visual imagery is famous in psychology. Their experiment about mental rotation showed that rotating an object mentally takes roughly the same time as it would take to rotate it in reality. Similarly, sport psychologists have known for a long time that "athletes can enhance their performance by merely mental rehearsal" (Goldman, unpublished, p. 12). Yue & Cole (1992) published a study in which they compared subjects who actually trained with subjects who generated appropriate motor imagery. The study shows that training leads to an increase of 30% in maximal force while motor imagery leads to an increase of 22%. This clearly shows that motor imagery has a great impact on strength, even if not as much as real training.

Goldman concludes that "all this indicates that pretense can often produce close facsimiles of naturally-generated states, which bodes well for the accuracy potential of pretense-based attributions of many types of states, including the attitudes" (p. 12). This suggests that pretend states won through simulation can in fact be used to make predictions and explanations of real states.

3.2 Simulation in More Detail

3.2.1 From First to Third Person Statements

The basic idea of simulation theory is that one uses one's own mental apparatus to simulate others and thereby comes to predictions and explanations. One motivation for simulation is that while the prediction of other people's behavior can be difficult, the prediction of "our own immediate and near immediate actions is usually a simple and accurate matter" (Davies & Stone, 1995b, p. 15). Gordon (1995a, p. 60) gives some samples of his own accurate self-predictions:

- I shall now pour some coffee.
- I shall now pick up the cup.
- I shall now drink the coffee.
- I shall now switch on the word processor.
- I shall now draft the opening paragraphs of a paper on folk psychology.

These statements on the whole are liable to be correct. Gordon notes that these self-predictions have "a success rate that would be the envy of any behavioral or neurobehavioral science" (p. 61). He then observes that these first-person statements are not very different to those of the third-person case. According to Gordon, the difference is only a matter of degree. Given that the first-person predictive statements are so accurate, one might wonder whether "the psychological mechanisms that are used in making them might be put at the service of more difficult predictive tasks" (Davies & Stone, 1995b, p. 16).

If one wonders how one would act if one were alone in the house and heard a sound in the basement, then one can use the same decision-making processes that one would use if one actually were in this situation in order to get an answer. The big question is whether it is possible to extend the first-person based methodology to third-person cases. Gordon (1995a) thinks that this is indeed possible:

> As in the case of hypothetical self-prediction, the methodology essentially involves *deciding what to do;* but, extended to people of 'minds' different from one's own, this is not the same as deciding *what I myself would do.* One tries to make *adjustments for relevant differences.* (p. 63)

In other words, you do not merely simulate that *you* are in the other's situation but the simulation is carried out having the other person's psychological traits. You therefore simulate *being the other*. When Robert M. Gordon (RMG) simulates someone, the referent "I" ceases to be RMG and becomes the person he is simulating. Gordon underlines this by saying that simulation requires not a transfer but a transformation (Davies & Stone, 1995a).

3.2.2 Some Examples of Simulation

Simulation theorists often use the example of the understanding of the effects of a new drug on the human body when talking about the advantages of simulation over a theoretical framework. When a new drug is developed and one wants to know which reactions it will have on the human body, the scientific knowledge of pharmacology and physiology could be used. This theoretical framework would allow us to generate hypotheses about the drug's effects. However, there is another method available to the researcher which is in fact commonly used in practice. One could

administer the drug to non-human animals which are similar to humans. In this case, one would, however, need a theoretical understanding in order to map the insights won in the experiments with non-human animals onto humans. However, when human guinea pigs are used, simulation theorists argue that this theoretical understanding is not necessary since "the bodily organs of human beings are being used in order to discover the effects of the drug" (Davies & Stone, 1995b, p. 17). It is important to note that in this example simulation is presented as an effective device; in this view, ST is not necessarily incompatible with TT. More radical simulation theorists would disagree with this point of view.

Another common example used to illustrate ST is the Tees/Crane experiment (Kahneman & Tversky, 1982). In the experiment, subjects are first read a story and then asked a question. The story is about Mr Crane and Mr Tees who are scheduled to leave the airport on different flights, at the same time. They go to the airport in the same car and get caught in a traffic jam, thus arriving at the airport 30 minutes after the scheduled departure time of their flights. Mr Crane is told that his flight left on time. Mr Tees on the other hand is told that his flight was delayed and left just five minutes ago. The question for the subjects then is: who is more upset?

The experiment showed that 96% subjects thought that Mr Tees would be more upset. Simulation theories usually claim that the subjects in the experiment come to an answer by simulating. They use their own mental apparatus to predict how the two characters in the story would feel. Of course they do not themselves become angry or resigned since their mental organs are operating off-line.

Finally, Fuller (1995) describes a story from his life to illustrate ST. He usually gets his sister a book for Christmas or her birthday and uses empathy and simulation to find the right book. Since he and his sister

have the same taste in literature, he simply reads the book himself and if he likes it he assumes that she will do so a well. According to Fuller, she has "rarely been disappointed" (p. 21).

3.2.3 Simulation Without the Concept of Belief

Gordon, who is one of the most radical simulation theorists, claims that simulation is sufficient for providing children with intentional concepts like belief, knowledge and desire. He says:

> 1. Let's do a Smith simulation. Ready? *Dewey won the election...*
> 2. *Smith believes that* Dewey won the election.
>
> My suggestion is that (2) be read as saying the same thing as (1), although less explicitly.
>
> (Gordon, 1995a, p. 68)

I will use the false belief task to illustrate this claim. In order to solve the problem of the false belief task, one has to – according to the simulation theorists – identify oneself with the person in the story (Sally or Maxi). Gordon says that one has to *imaginatively identify* with Sally and imagine the world from Sally's point of view. Even though the child knows that the marble is not in the basket, she (from Sally's point of view) has to hold the following thought:

> I [Sally] believe that the marble is in the basket.

Gordon then argues that one does not need to have a concept of belief at all in order to hold this belief. The "I believe that" could just as well be deleted. The child, who identifies with Sally, thus simply has to hold the thought

> The marble is in the basket.

The child does not need the *full concept of belief* – it merely needs to be able to have *beliefs*!

In order words, according to Gordon

I believe that p

can always be reduced simply to

p

as seen in his example with Smith and Dewey given at the beginning of this section.

This claim of Gordon is quite controversial among theory theorists as well as simulation theorists. Perner & Howes (1992), who evaluate different ST positions, describe a more cautious claim of simulation theorists which suggests that simulation is "at best necessary (but not sufficient) for the acquisition of intentional concepts" (p. 75). In this view, simulation provides a useful, and perhaps necessary, source of data for the acquisition of concepts but "does not in itself constitute understanding of these concepts" (p. 75).

The question of the relation between simulation and the concept of intentional states and the understanding of psychology is answered differently by many simulation theorists. Gordon's claim is certainly the most interesting one but at the same time the most controversial and criticized one. I will discuss the various positions in more detail later.

3.2.4 ST and the False Belief Task

As seen in chapter 2.4, the false belief task is commonly used to test for the presence of a theory of mind which is comparable to that of adults in the main respects. While it has been shown how theory theorists

explain the failure of children younger than four in the false belief task, the explanation of the simulation theorists has not been given yet.

Simulation theorists obviously think that five year old children solve the problem through simulation. Therefore, the simulation has to fail for younger people. The question is "why". Harris suggests that younger children fail because the complexity of the imaginative task that they face is beyond them. The simulation would require "overwriting current reality (in which the marble is in the box), and adopting the divergent stance towards the world that imaginatively taking on that belief involves" (Davies & Stone, 1995b, p. 31). Davies & Stone conclude that:

> The child of five succeeds where the child of three fails because the older child has an ability to simulate another whose view upon the world is different from his own. (The child subject saw the marble moved from the basket to the box; Sally in contrast was out of the room, and so quite literally had a different view.) (p. 6)

This view contains an important implication. While the implication is not surprising, it has not yet been expressed explicitly in the discussion of simulation. The explanation of the false belief test suggests that simulation is an ability rather than knowledge. While development has been explained as theory reductions and changes in knowledge by the theory theorists, simulation theorists view development as the refinement of a skill. Children gradually become "more adept at imaginatively identifying with other people and at imaging counterfactual situations" (Davies & Stone, 1995b, p. 6).

3.2.5 ST and Theoretical Knowledge

Although the view that simulation heavily depends on the refinement of a skill does not seem surprising, the question whether simulation is merely a skill or whether it requires skill and knowledge is not uncontroversial, as seen before in the discussion of the relation between simulation and the concept of intentional states. It is clear that any simulation requires *information*. When I want to simulate someone else being alone at home, it helps to know whether the person is of a frightened and scared nature. However, while this knowledge is important for a particular simulation, it is not substantial for simulation *per se*. The question is whether simulation requires the knowledge of a theory in order to function at all. Perner (1991) argues that no simulation can do without theoretical knowledge:

> Assume you learn that your colleague's mother-in-law has just died. How does he feel? It ill not do to imagine that your mother-in-law has just died [...] because your relationship [...] may be quite different from your colleague's. [...] Your simulation must be informed by some "theory" about which personal relationships are emotionally relevant. If you love your mother-in-law but your colleague hates his, then your simulation will be more accurate if you imagine the death of one of your foes. (p. 268)

Gordon (1995b) accepts that simulation needs evidence but questions if it has to be in the form of a theory. He wonders how one uses evidence in order to come to a prediction and suggests that it happens "not by plugging the evidence into a general theory of the organization of the human behavior control system, but by trying to motivate similar behavior within the context of simulation" (Gordon, 1995b, p. 110).

3.3 Support for ST from Contemporary Research

Simulation theory has gained a considerable momentum during the last few years. This is in part due to new insights won in elaborate research carried out in areas which were not connected to theory of mind in the past – the new insights, however, showed that there are more connections than previously thought. I am going to describe research results from two fields which seem to support the simulation view of theory of mind.

3.3.1 Autism

The biologically based disorder of autism is often described by simulation theorists as supporting ST. However, this claim is quite controversial. Although an impairment of theory of mind is increasingly used by autism researchers to explain the disorder, some prominent researchers (such as Alan M. Leslie) use TT approaches instead of the simulation view.

Currie (1996), who examines ST, TT and the evidence from autism, describes the two theories as the ability of *knowing that* (TT) and *knowing how* (ST). Instead of saying what TT and ST are he lists what people with autism *lack* according to the theory. Theory theory suggests that autism is a deficiency of knowledge. According to this theory, autistic people do not know a significant number of propositions and perhaps cannot even formulate them. ST says that "autism is a deficiency of imaginative capacity – the capacity to project the self imaginatively into a situation other than its own current, actual position" (Currie, 1996, p. 243).

The reason why many simulation theorists use autism as a case supporting ST is because of two notable aspects of autism: deficits in pretense (and imaginative activity in general) and mentalistic understand-

ing. The latter is clearly due to an impairment of theory of mind and the former shows many connections to simulation, especially to the off-line variant. In simulation, you have to pretend being the other person. However, imagination and pretend play is obviously not possible without being able to "separate real states of affairs from states of affairs that are being pretended about" (Frith, 1996, p. 65). Without this, simulation cannot work since you can hardly feed your own decision-making process with pretend inputs such as pretend beliefs and pretend desires without having the capability of pretense.

What is striking is that autistic people (even after the age of four) have great difficulty with false belief tests. The Sally experiment has been carried out with autistic patients and the results showed that most subjects did not pass the test. ST explains the failure in terms of simulative incompetence. The autistic person does not understand that the puppet Sally "will look in the wrong place for her sweet because she has an impaired capacity to simulate Sally, whose epistemic situation differs from the child's own in so far as she lacks knowledge the child possesses" (Currie, 1996, p. 248).

Similarly, autistic people fail in the Smarties false belief task. The child is shown a closed Smarties tube and has to guess the contents. Naturally, most children give "Smarties" or "sweets" as a reply. When the tube is opened, a pencil comes out to the child's suprise. The next question is what Billy, who will arrive soon, will say when asked about the contents of the tube. Most children say that he will give "pencil" as an answer although they are aware what they first thought to be the contents before seeing the pencil (Baron-Cohen, 1995, p. 71).

However, it is to note that while most autistic people fail these tasks, there are some who pass (about 20%). This variability of autistic performance is quite hard to explain using TT since people using the same

theoretical framework would reach the same conclusions. However, ST, which is seen as an ability, has two explanations to offer. First, some people with autism have learned to "hack out" a solution for these rather simple and artificial tests but they do not really have the ability to mentalize. This implies that success on false belief tasks is only to be found in older autistic children "who have learned various rules of thumb, for example that people who haven't seen something don't know about it" (Currie, 1996, p. 247). Second, different autistic people have a varying degree of difficulty with simulation and those less affected can make simple perspective shifts required for the false belief tests.

Both explanations imply that autistic people will fail in more complex false belief tasks, which – it has been shown – in fact they do. The first explanation fits neatly with the suggestion of Goldman (unpublished) that evolution gave us more than only one mentalizing strategy (see page 25). Since their simulation ability is impaired, people with autism will try to acquire a theoretical framework in order to master their basic life. This suggestion does indeed not seem unlikely since autistic people have been shown to be very good with technical things and at theorizing. This explanation, however, is not compatible with Gordon's point of view that simulation is necessary for the formation of mental concepts.

It is not clear yet how much of autism can be explained by an impairment of theory of mind. There are more aspects of autism than just impairment in play, social interaction, and verbal and non–verbal communication. Happé (1999) argues that all "deficit accounts of autism [...] fail to explain why people with autism show not only preserved but also superior skills in certain areas" (p. 217). This should not concern us, however, since it is plausible that the impairment of theory of mind due to a simulative incompetence is still true – it just has to be

supplemented by another explanation. The case of autism is thus a good example supporting ST.

3.3.2 Mirror Neurons

A new type of visuomotor neuron which has recently been discovered in the monkey's premotor cortex has gained significant popularity within cognitive science. It has been suggested that these neurons are used for imitation, allow the acquisition of language (Rizzolatti & Arbib, 1998) and enable theory of mind (Gallese & Goldman, 1998).

Neurons which discharge when the monkey grasps or manipulates objects have been known for a long time. Recently, a new type has been found which also discharges when the monkey observes the experimenter making a similar gesture. Rizzolatti & Arbib (1998) describe these neurons, dubbed "mirror neurons" (MN), in more detail:

> The response properties of mirror neurons to visual stimuli can be summarized as follows: mirror neurons do not discharge in response to object presentation; in order to be triggered they require a specific observed action. The majority of them respond selectively when the monkey observes one type of action (such as grasping). Some are highly specific, coding not only the action aim, but also how that action is executed. They fire, for example, during observation of grasping movements, but only when the object is grasped with the index finger and the thumb. (p. 188)

Gallese & Goldman (1998) suggest that one function of these neurons is to "enable an organism to detect certain mental states of observed conspecifics. This function might be part of, or a precursor to, a more general mind-reading ability" (p. 493). They argue that mind-reading could make a contribution to inclusive fitness since "detecting another

agent's goals and/or inner states can be useful to an observer because it helps him anticipate the agent's future actions, which might be cooperative, non-cooperative, or even threatening" (p. 495–496).

They also claim that simulation can be used to retrodict as well as predict mental states. In other words, it is possible to determine which mental states of a target have *already* occurred. The attributor can wonder what goals the target had that led him to perform an action. He can then go backwards and draw an inference from the observed action to a hypothesized goal state. This mechanism can be used to generate explanations of the target's behavior. Furthermore, the mirror neuron system can be used to produce the target's mental states in the attributor since the mirror neurons are active when observing another monkey making an action. This implies that there is one system with two distinct (but closely related) functions:

> MNs respond both when a particular action is performed by the recorded monkey and when the same action performed by another individual is observed. All MNs [...] discharge during specific goal-related motor acts. Grasping, manipulating and holding objects are by far the most effective actions triggering their response. (Gallese & Goldman, 1998, p. 495)

In summary, NM activity "seems to be nature's way of getting the observer into the same 'mental shoes' as the target – exactly what the conjectured simulation heuristic aims to do" (Gallese & Goldman, 1998, p. 497–498).

3.3.3 Combining Autism and Mirror Neurons

One might wonder if there are any connections between autism and mirror neurons. After all, the ability to detect and represent mental states of others seems to be impaired in autistic people and this is – so it has been suggested – one function of mirror neurons.

Castelli et al. (2000) have conducted a study in order to find out which brain areas are involved in theory of mind. They used the PET (positron emission tomography) technique to get images of the brain activity of six healthy adult volunteers. The volunteers were watching computer-presented animations which generally lead to the attribution of mental states. The study clearly showed that some specific areas were more active during the mentalizing tasks than during the control period. Abell et al. (200) went one step further and showed that these brain areas are less active in autistic children.

It has been suggested that the brain areas which have been found in these studies correspond to the areas where mirror neurons are generally found. It is not clear yet whether mirror neurons specifically or any type of neurons found in these particular areas are involved. It would not be surprising, however, if evidence was found which showed that mirror neurons played a central role. More research has to be carried out in order to shed more light on these issues, but it certainly is intriguing that there might be connections between two areas which seem to support ST. It would strengthen the hypothesis of ST enormously if the case of autism, mirror neurons and ST could be combined to form better explanations of all issues involved.

3.4 Different ST Positions in More Detail

Similar to theory theory, of which many different versions have been proposed, simulation theory is not one coherent theory. Instead, the term "simulation theory" is used as an umbrella term to refer to different theories which have in common that they view mindreading as an ability rather than the use of a theoretical framework. In the following, I am going to give an overview of three of the most popular versions of simulation theory. I will discuss the views of Paul L. Harris, Alvin I. Goldman and Robert M. Gordon.

3.4.1 Harris

Harris' approach to understanding theory of mind is one of developmental psychology. He sets out the "psychological case, and more specifically the developmental case, for the proposal that children improve their grasp of folk psychology[2] by means of a simulation process" (Harris, 1995, p. 207). He maintains that improvement in the mindreading ability stems "from changes in the child's imaginative flexibility, rather than from a transformation in the child's so called theory of mind" (p. 216). In order to execute appropriate simulation thoughts, those acts must usually override a "background of default settings" (Harris, cited in Davies & Stone, 1995b, p. 30). Children are increasingly able to deal with the high number of adjustments needed in order to perform accurate simulations.

Harris' approach to simulation does not claim that simulation is always accurate – not even in adults. The accuracy depends, according to Harris (1995, p. 226):

[2]Harris uses the term "folk psychology" in a different way than I do. He uses it to refer to the general mindreading ability or understanding of theory of mind, not to the theoretical framework postulated by theory theory.

1. on feeding in the relevant pretend inputs, and

2. on the target behavior being guided by the decision-making system.

If these two assumptions are not met, inaccurate predictions are more than likely. Harris, who is mostly concerned with the problem how children acquire and improve their mindreading ability, summarizes that:

> In the course of development, children become increasingly proficient at feeding in the appropriate pretend inputs. Much of that advance is constrained by increments in imaginative power. (p. 226)

In his discussion about theory theory and simulation theory, Harris gives the following thought experiment: English speakers are presented with grammatical and ungrammatical sentences and they have to make judgments about which is which. Another person is then asked to predict what decision most people came to about each sentence. It turns out that the person's hit rate is very high.

> In almost every case you can tell me whether the majority judged the sentence to be grammatical or ungrammatical. Moreover, when I ask you to explain your predictions you do so by indicating deviant constructions or morphemes in the ungrammatical sentences, something that speakers in the first part of the study also did. How are your predictions so accurate? (Harris, 1995, p. 210)

Harris replies that the most plausible answer is that one simply read each sentence and asked oneself whether it sounded grammatical or not. Furthermore, one assumed that other English speakers would come to the same conclusions for the same reason.

Stich & Nichols (1995) agree that the sort of predictive strategy Harris sketches generalizes to many other cases. They wonder what their

Rutgers colleagues would say when asked: "Who is the President of Rutgers University?" and conclude that "we would proceed by first answering the question ourselves [...] and since we assume that our colleagues [...] believe the same things we do on questions like this, we would predict that they would say the same thing we would" (p. 93).

While Stich & Nichols agree that this strategy can be seen as a sort of simulation, it is radically different to off-line simulation, which is so typical for ST. They then try to explain the difference:

> On the off-line simulation account, the information about the target's belief (along, perhaps, with some other information about the target's desires and further beliefs) is fed into our own decision-making system. That makes a decision which, rather than being acted on, is transformed into a prediction and reported to the "belief box". (p. 94)

In the thought experiment about grammatical and ungrammatical sentences, however, simulation does not take the states of the target into account. One's decision-making system is not fed with such states either. Instead, we first determine what we believe ourselves and then attribute the same belief to someone else. While Stich & Nichols agree that this is a very plausible account of the way in which we sometimes figure out what other people believe, they are quite sceptic to what amount this type of simulation contributes to ST. They observe that:

> It is *not* a process which results in predictions of behavior, and it is *compatible with* both the theory-theory and the off-line simulation theory, each of which provides an account of how we use information about a person's beliefs, desires, and other mental states in producing predictions about that person's behavior. (p. 94)

Stich & Nichols dub this type of simulation "type-1 Harris simulation" and go on to sketch another, more controversial sort of simulation,

which they believe Harris has also in mind. They begin with a "Harris-style thought experiment" (p. 94):

> Sven believes that all Italians like pasta. Sven is introduced to Maria, and he is told that she is Italian. (p. 95)

Now someone is asked what Sven would say if he was asked: "Does Maria like pasta?" One way to answer this question is to make use of a theoretical framework about how people form beliefs from other beliefs. Using theory theory, one could infer that Sven will come to believe that Maria likes pasta. Another way to solve the problem would be to use a simulation process and "feed pretend or hypothetical inputs into your own inference mechanism, and then allow it to churn away as it normally does and draw appropriate conclusions" (Stich & Nichols, 1995, p. 95). The pretend inputs would be:

> All Italians like pasta.

and

> Maria is an Italian.

Quite obviously, the conclusion would be:

> Maria likes pasta.

According to Stich & Nichols, the conclusion that Maria likes pasta is not directly given to your belief box. If it were, *you* would end up believing that Maria likes pasta, rather than you believing that Sven holds this belief. Therefore, the conclusion "churned out by your inference mechanism is attributed to Sven" (p. 95). This implies that there must be a mechanism which takes "the output of your inference box and embeds it in a belief-sentence before it is fed into your belief box" (p. 95). Stich & Nichols call this inference simulation process "type-2 Harris simulation".

The most interesting insight won through the discussions of Stich & Nichols besides the detailed distinction between type-1 and type-2 Harris simulation is that both forms seem to be compatible with theory theory. While Harris (1995) maintains that children learn to grasp mindreading through simulation, he agrees that "there is no reason to doubt that adults resort to theories, be they tacit or explicit, in explaining and predicting behaviour" (p. 226). Therefore, it would not be surprising if he largely agreed with the observations made by Stich & Nichols. In an "one-eyed overview of a debate", Leslie & German (1995) give their summary of Harris' version of simulation theory:

> Harris has in mind a notion of simulation that is very broad indeed, encompassing almost any use of one's own knowledge in the interpretation of another person's behavior, including, for example, using one's knowledge of English to understand what someone says to you. This will almost guarantee that most theory of mind abilities involve "simulation", but such an outcome is largely a terminological victory. (p. 127)

This problem becomes very obvious in the example of Sven and Maria. The pretend inputs of the simulation are typical premises used in logical deductions. If it is held that a simulation is used instead of a deduction, then the boundaries between simulation and TT begin to blur.

3.4.2 Goldman

In his important 1989 paper "Interpretation Psychologized" in Mind and Language,[3] Alvin I. Goldman asks how people arrive at attributions of propositional attitudes (and other mental states). In his investigation, he assumes that the interpreters themselves have beliefs and finds it hard to imagine the problem without this assumption. He points out that there are three types of interpretation theories:

1. Rationality theories which rely on the basic presumption that an "attributor A operates on the assumption that the agent in question, S, is rational" (Goldman, 1995a, p. 75). In other words, the agent acts according to an ideal or normative model of proper inference and choice. Goldman names Daniel C. Dennett's intentional stance as an example for a rationality theory. Intentional states are attributed using the intentional stance by "first postulating ideal rationality on the part of the target system, and then trying to predict and/or explain the system's behavior in terms of such rationality" (p. 76).

2. Folk theory theories which postulate that attributors somehow acquire a common-sense or folk psychological theory. Goldman claims that such theories face three problems: vagueness, inaccuracy, and non-universality. First, the laws constituting the folk psychological theory are so vague that no reliable interpretive conclusion can be drawn. Second, the theory is only useful if the laws are actually more or less true. However, Goldman doubts that "ordinary interpreters [...] posses laws that are true" (p. 79). Third, a common assumption of theory theories is that most competent users of the

[3] Reprinted as Goldman (1995a).

folk psychological theory share a common set of laws and platitudes. Goldman thinks that "this universality assumption is very dubious" (p. 79).

3. Simulation theory, which Goldman describes and defends in his 1989 and other papers, such as Goldman (1995b).

Goldman's view of simulation theory is a very interesting account of simulation. Goldman (1995a) suggests that we do not use a mathematical decision theory to make predictions but rather "consider what *we* should do if we had the relevant beliefs and desires" (p. 81). One ascribes mental states to others by pretending and imagining oneself being in the other's shoes. One first generates the states in which the person is in oneself and then acts thereupon. Goldman summarizes this process by saying that "we *simulate* the situation of others, and interpret them accordingly" (p. 81). While simulation is often described as an effective heuristic to predict beliefs, desires and other states, it is "also relevant in inferring *actions* from mental states, not just mental states from other mental states" (p. 82).

It is important to note an important difference to Gordon's view of simulation (which will be covered right after Goldman). While Goldman claims that we put *ourselves* in the shoes of someone else and simulate the other person, Gordon sees simulation as a transformation rather than a transfer. In other words, during simulation Robert M. Gordon (RMG) ceases to be RMG and instead becomes the person he simulates. Such a transformation is not postulated by Goldman's theory, however.

In his introduction to simulation, Goldman suggests that it seems – introspectively – as if we were using simulation quite regularly in order to try to predict the behavior of the people around us. We often imagine ourselves in their situation and determine what they might think. For ex-

ample, when we are playing chess, I may try to predict the other player's next move by imaging myself being in their position and deciding what I would choose to do. Goldman (1995a) summarizes this method and its assumptions:

> From your perceptual situation, I infer that you have certain perceptual experiences or beliefs, the same ones I would have in your situation. I may also assume (pending information to the contrary) that you have the same basic likings that I have: for food, love, warmth, and so on. (p. 82)

Goldman underlines that the simulation procedure cannot be used too simplistically. In order to come to adequate predictions using simulation, one has to pay attention to individual differences. Goldman suggests that if "I am a chess novice and you are a master, or vice versa, it would be foolish to assume that your analysis would match mine" (p. 82). In order to optimize the use of simulation, one must not only imagine oneself being in possession of the other's goals and beliefs but also in possession of the other's level of chess sophistication. Goldman notes that people may not always take such factors into account, or frequently lack the information needed to make an adequate and accurate adjustment. He does therefore not assume that people are always successful or optimal simulators. As such, he does not propose that simulation is a perfect method or the only method used for interpersonal mental ascriptions or for prediction of behavior. He proposes, however, that simulation is "the fundamental method used for arriving at mental ascriptions of others" (p. 83).

While Goldman views simulation as the primary method of forming mental ascriptions and predicting other people, he suggests that it is very plausible that evolution gave us more than only one mentalizing strategy (Goldman (unpublished); also see page 25 where I have discussed

this before). In particular, he names theory theory as an alternative to simulation. However, he also mentions that the use of a theoretical framework alone is not as effective as simulation, as seen in the cases of autistic people who are able to complete easy false belief tasks, but who largely fail on harder tasks and in real world situations.

In fact, Goldman (1995a) claims that we often develop "generalizations and other inductively formed representations (schemas, scripts, and so forth) that can trigger analogous interpretations by application of [...] 'knowledge structures' alone, *sans* simulation" (p. 88). As an illustration, Goldman gives the example of Jones and Brown. Jones always greets people with a smile whereas Brown greets them with a grunt. People can then form expectations without the use of simulation since there is a regularity in the behavior of Jones and Brown. As a second example, Goldman says that we do not need simulation in order to predict that people who enter a car in the driver's seat will typically proceed to start it. Goldman views simulation as an intensively used heuristic on which interpretation fundamentally rests, but accepts that inductive or nomological information is not wholly absent. However, it plays a less important role than in theory theory. Goldman (1995a) summarizes that "simulation remains the fundamental source of interpretation, though not the essence of every act (or even most acts) of interpretation" (p. 88).

Additionally, Goldman suggests that we use simulation to a larger extend that we think. While we know that we sometimes use simulation to predict other people's behavior, we are typically not aware of simulation processes going on in us. Goldman (1995a) suggests that this is because simulation need not be an introspectively vivid affair and because it is likely that the process is "semi-automatic, with relatively little salient phenomenology" (p. 88). He explains that

> [i]t is a psychological commonplace that highly developed skills become automatized, and there is no reason why interpersonal simulation should not share this characteristic. (On the issue of conscious awareness, the simulation theory is no worse off than its competitors. Neither the rationality approach nor the folk-theory theory is at all credible if it claims that appeals to its putative principles are introspectively prominent aspects of interpretation.) (p. 88)

Furthermore, as mentioned above, simulation is an effective heuristic which can be used to develop generalizations. Therefore, simulation has to be used in an decreasing amount since more and more generalizations are available (and which are – once developed – typically easier and faster to use than carrying out a simulation process).

One notable aspect of Goldman's view of ST is that he distinguishes between two types of simulation. In response to Dennett's question how simulation can work without being a kind of theorizing, Goldman (1995a) claims that there are two different variations of simulation. If a person wants to simulate the weather or the economy successfully and accurately, they will fail unless they have a good theory of the system. This is, according to Goldman, a *theory-driven* simulation. The question remains whether *all* simulations are theory-driven. Goldman negates this question and proposes *process-driven* simulations as an alternative. This alternative is only possible if two conditions are met:

1. the *process* that drives the simulation is the same as (or relevantly similar to) the process that drives the system, and

2. the initial states of the simulating agent are the same as, or relevantly similar to, those of the target system.

   (Goldman, 1995a, p. 85)

A person who tries to simulate a sequence of mental states of another person will wind up in the same final states if they begin in the same initial states and if both sequences are driven by the same cognitive process or routine. While these two requirements have to be met, it is of no importance at all for the simulating agent to have "a theory of what the routine is, or how it works" (p. 85). Process-driven simulation is therefore an effective means to predict and explain other people's beliefs, desires and states if the two conditions are met.

In fact, there is another assumption which underlies process-driven simulation. I have already discussed this precondition briefly in chapter 3.1 when I covered AD (attributor dependent) heuristics, for which simulation is the prime example. The first condition of process-driven simulation listed above is that the interpreter needs to be in the very same initial states as the interpretee. However, it is not possible to be in exactly the same states. While they might share some desires and goals, there will always be relevant differences, too. One possible way to get around this problem is to imagine or feign the same initial states as the interpretee has. However, this raises the question whether these *pretend* states are "relevantly similar to the genuine beliefs and desires that they model" (Goldman, 1995a, p. 85). Goldman (unpublished) argues quite successfully that they are in fact sufficiently similar, as seen in evidence from various domains. For an overview, refer to chapter 3.1, especially the discussion on page 25.

An interesting implication of Goldman's distinction between real and pretend beliefs and states is that it seems as if he had the assumption of two different systems. This is in stark contrast to Gordon, who clearly postulates one system. It is not clear how this distinction fits in with the evidence won in research about mirror neurons (in which Goldman was involved). Mirror neurons respond both when a specific action is

performed by a recorded monkey and when the same action carried out by another individual is observed (Gallese & Goldman, 1998). This implies one system since the same mirror neurons are active in both cases. Unfortunately, Goldman does not clearly state how the mirror neuron system is related to his distinction between real and pretend states.

Summing up Goldman's view of ST, it can be said that his account of simulation is a very interesting and promising approach. He sees simulation as a very effective heuristic, but also accepts that inductive or nomological information is not wholly absent. This makes Goldman's account a "less-than-radical, knowledge-*and*-ability account, where one of the abilities happens to be simulation" (Leslie & German, 1995, p. 132). However, simulation remains the fundamental source of interpretation and it is where "the action is" (Goldman, cited in Leslie & German, 1995). Also, while his view of simulation converges somewhat toward theory theory, ST clearly remains distinct:

> It still remains distinct, however, (A) because the folk-theory theory makes no allowance for simulation, and (B) because the complex variant postulates simulation as the originating source of (most) interpretation. (Goldman, 1995a, p. 88)

Another feature of Goldman's "less-than-radical" account is that he is not very optimistic that simulation will yield a "constitutive account either of mental states or of the possession of conditions for mental concepts" (Davies & Stone, 1995a, p. 5). While simulation can be used in order to form generalizations, it cannot be necessary for the formation of mental concepts since Goldman suggests that autistic people use theories as an alternative strategy. This view represents a huge difference to Gordon's position which claims that simulation will indeed yield an alternative to the theory theorist's understanding of mental states.

3.4.3 Gordon

Gordon's version of simulation theory is the most controversial view of simulation which has been proposed to this date. It is so different to other theories because simulation plays a radical different epistemological significance for Gordon than suggested in other theories:

> That people do sometimes resort to [...] simulation is not in serious dispute. What is in dispute is the claim that simulation is fundamental to [theory of mind] or at least is of deep psychological and philosophical significance. (Gordon, 1995c, p. 53)

Most simulation theorists view simulation as an effective heuristic for predicting the behavior of others. Gordon, however, goes one step further and holds that, beyond this, "even our ability to grasp the concepts of mind and the various mental states depends on our having the capacity to simulate others" (Gordon, 1996, p. 11). This claim is the reason why Gordon's version of ST is often called the "radical" position.

Gordon (1995c) describes "traditional" simulation theories and argues that they depend on an implicit inference from oneself to others, not unlike the use of an analogy. This is usually connected to introspection and the question of "*how* one recognizes and ascribes one's own mental states" (p. 53). He summarizes that, according to this account, simulation is:

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 (1995c, p. 53)

Gordon continues by stating that he rejects all of these three assumptions. Additionally, he claims that most arguments against the simulation theory "crucially depend on the assumption that in simulating another one recognizes one's own mental states by introspection and then infers that the other is in similar states" (p. 54). Since Gordon's theory does not depend on introspection and inference, he maintains that the arguments are mute when applied to his theory.

Gordon (1995c) uses the Tees/Crane example to illustrate his version of simulation. In order to find out what Mr Tees would think, Gordon does not imagine what he would do in Tees' situation. Instead, uses an alternative way to solve the problem: "I have the option of imaging in the first person *Mr Tees* barely missing his flight, rather than imaging *myself*, a particular individual distinct from Mr Tees, in such a situation and then extrapolating to Mr Tees" (p. 55). RMG then ceases to be the referent of "I" and "I" refers to Mr Tees instead. This is due to the egocentric shift which is required by Gordon's simulation. This leads to Gordon's slogan that simulation is "not a transfer but a transformation" (Gordon, 1995c, p. 54). This view leads to the implication that neither introspection nor an inference are required:

> The point I am making is that once a personal *transformation* has been accomplished, there is no remaining task of mentally *transferring* a state from one person to another, no question of *comparing* Mr Tees to myself. For insofar as I have recentered my egocentric map on Mr Tees, I am not considering what *RMG* would do, think, want, and feel in the situation. Within the context of the simulation, RMG is out of the picture altogether. In short, when I simulate Mr Tees missing his flight, I am already representing *him* as having been in a certain state of mind.

In order for the transformation to work properly, one has to decide what mental states Mr Tees is in. For example, we have to determine "whether he was *extremely upset*, whether he *thought it was the driver's fault*, and so forth" (Gordon, 1995c, p. 57). Critics could claim that the capacity for introspection is needed to solve this problem. Gordon, however, has a different alternative to offer:

> My own view [...] is that the method we ordinarily use is limited to identifying states in the *first person*, but, thanks to our capacity for imaginatively transforming ourselves into *other* "first persons", it is not exclusively a *one-person* method. It is just as well suited for labeling another's states as it is to labeling our own, provided we represent these states in the first person, that is, by an egocentric shift. (p. 58)

This egocentric shift is possible due to a technique which Gordon (1996) calls "ascent routines". These routines have two purposes: they are used for self-ascriptions as well as for egocentric shifts.

Gordon thinks that adults often answer questions about their belief p by asking themselves the question whether or not p is true. For example, if someone asked Gordon "Do you believe Mickey Mouse has a tail?" ($Q1$) he would ask himself "Does Mickey Mouse have a tail?" ($Q2$). If the answer for $Q2$ was "yes", he would give the same answer to $Q1$. If he would negate $Q2$, he would do the same with $Q1$.

> I call this procedure an *ascent routine* because it answers a question by answering another question pitched at a lower semantic level – the former being a question about a mental state that is about x, the latter a question directly about x. (Gordon, 1996, p. 15)

This procedure has an interesting implication:

> What is of particular interest is that it allows one to get the answer to a question about oneself, and specifically about one's mental states, *by answering a question that is not about oneself, nor about mental states at all.* (Gordon, 1996, p. 15)

One implication from this feature of ascent routines is that they are equally well suited to identifying another's beliefs as it is to identifying one's own, as has been mentioned briefly before. If you want to answer "Do I believe that p?" during the simulation of a person O, you can simply look at the situation and determine whether it is the case that p. In Gordon's words:

> So I settle the question of whether O beliefs that p simply by asking, within the context of a simulation of O, whether it is the case that p. That is, I simply concern myself with *the world* – O's world, the world from O's perspective [...] – and, reporting what is there, I am reporting O's beliefs. That is, reporting O's beliefs *is* just reporting what is there. (Gordon, 1995c, p. 60)

Another feature of ascent routines is that they give children the capability to answer questions they would normally not be able to answer since they transfer a difficult question (about your mental states) in an easy question (about the world). However, this does not actually equip them with genuine, comprehending ascriptions of belief. They would not understand that the question "Do you believe that p?" is a "question about *themselves* rather than simply a question *about* (for example) *Mickey Mouse*" (Gordon, 1996, p. 16). Gordon summarizes this as follows:

> The point is [...] that they would have no means of understanding how, 'I believe Mickey Mouse has a tail', could be

> about an individual (other than Mickey Mouse) at all. They fail to grasp several components of the concept of belief, but the one that is paramount, because it is presupposed by all the others, is the general idea that a fact (about Mickey Mouse, for example) can have a *mental location*: can be, in other words, a fact *to* some individual. (p. 16)

While ascent routines do not equip children with a genuine understanding of mental concepts, Gordon maintains that simulation is used to bootstrap such an understanding. This obviously implies that simulation has to work without understanding mental concepts. Gordon (1995b) gives an example to illustrate that this is possible: "Long before the child is able to attribute to herself or another an *interest* in something, she will turn her eyes to what the other is gazing at; and at a later stage, pull up alongside another child who is studying an object on the floor" (p. 114). The question how children actually develop a true understanding of mental concepts remains, however.

According to Gordon, children may realize by using their ability to simulate that "assertions within the context of a simulation can contradict [their] own (unpretended) beliefs" (Davies & Stone, 1995a, p. 13). A child will not learn to understand that its beliefs may deviate from the facts by simply asking what the facts are.

> To see her own present beliefs as distinguishable from the facts she will have to simulate another for whom the facts are different – or, more broadly, adopt a perspective from which the facts are different, whether this perspective is occupied by a real person or not – and then from the alien perspective, *simulate herself*. (Gordon, 1995c, p. 62)

This is the first step in the direction of genuine understanding of the notion of belief. The child will realize abstractly that her present

perspective which she views as fact may indeed not be fact at all but "nothing more than fact-from-her-particular-perspective" (p. 62).

Based on these observations, Gordon (1995c) distinguishes between two types of ascriptions and draws his conclusions about their development:

> [If] we ordinarily identify our own present beliefs by using an ascent routine, then there is an important distinction to be made between *comprehending* and *uncomprehending* ascriptions: that is, ascriptions made with and ascriptions without the understanding that the beliefs ascribed may be false. On the one hand, a capacity for reliable *uncomprehending* identification of one's own present beliefs should emerge before one can ascribe beliefs to others or to oneself in the past. It emerges extremely quickly, if my view is right, and *does not even await development of a capacity to introspect*, much less a capacity to recognize a belief by its introspected phenomenological marks. (Gordon, 1995c, p. 62)

Gordon's "radical" version of simulation theory is quite different to Harris' or Goldman's versions described before. Gordon argues that Harris' and Goldman's theories depend on introspection and inference while his own theory suggests a transformation. Gordon's ascent routines should do away with the need of the problematic introspective access. Furthermore, they offer interesting perspectives of how children and adults use and learn beliefs. Only by simulating, children can recognize that what they hold as "facts" are only beliefs. Through the use of simulation, they eventually acquire a genuine understanding of mental concepts. This is in stark contrast to Harris and Goldman who do not believe that simulation can lead to the formation of mental concepts.

3.5 Problems of ST

In recent years, simulation theory has brought new momentum to the philosophy of mind debate. The new position is certainly an interesting view and should be considered and evaluated thoroughly. Since ST has gained importance in the philosophy of mind debate, many objections have been found and described. I will present some of the most important arguments against ST in the following.

3.5.1 Simulation Needs a Theory

Several arguments have been put forward which claim that simulation requires a theory so that it can be used successfully. The arguments vary in the degree in which they allow simulation to work without the use of a theory. While some critics have argued that simulation can indeed work without a theory, but that you need one in order to form explanations, others believe that a theory is already required to carry out a simple simulation. In the following, I will discuss three different arguments which suggest that simulation has to be combined with a theory.

Dennett (1987) wondered how simulation can work without being a kind of theorizing. He suggested that he was a suspension bridge and observed how a simulation could be carried out:

> If I make believe I am a suspension bridge and wonder what I will do when the wind blows, what 'comes to me' in my make believe state depends on how sophisticated my knowledge is of the physics and engineering of suspension bridges. (p. 100)

These observations made Goldman (1995a) propose a distinction between theory-driven and process-driven simulation. While Goldman admits that theory-driven simulation is in need for a theory, this is not the case for process-driven simulations. Process-driven simulation makes use

of the fact that your decision-making apparatus is very similar to the person you are simulating. You therefore do not need a theory how the mechanism works. You simply feed your apparatus with input data and let it do the work.

Another argument put forward is that you need a theory in order to get started with simulation. This view has been summarized by Davies & Stone (1995b) who write that "even if mental simulation does not need to be driven by a psychological theory, still theory comes in when we try to set the simulation up in the first place" (p. 19). When you perform a simulation, you have to feed your decision-making apparatus with appropriate pretend inputs. However, the argument goes, you need a theory in order to determine which facts are really relevant to the simulation:

> [T]aking those variables into account is a matter of reflecting upon a number of theoretical considerations. [...]
>
> [W]hen we simulate another person, we need to make allowances for relevant differences. But which differences are relevant? (Davies & Stone, 1995b, p. 19)

The simulation theorists have a good defense for this argument, however. They can claim that you do not need a theory at all. You simply put yourself in the shoes of the person you wish to simulate and let your decision-making apparatus do the rest. It will select the information it requires for its decision and will discard the rest. If not enough information have been supplied, the outcome of the simulation will not be reliable. However, simulation theorists openly admit that simulations are not always perfect and heavily depend on having enough information supplied.

The final argument is that while simulation can be used to predict other people's behavior, a theory is needed to form *explanations*. Churchland (1991) argues:

> A simulation itself, even a successful one, provides no explanation. What explanatory understanding requires is an appreciation of the *general patterns* that comprehend the individual events in both cases. And that brings us back to the idea of a moderately general *theory*. (p. 60)

Simulation theorists usually reply that simulations can be used to test the behavior under different circumstances and thereby find the causal factors. For example, Gordon (1995b) writes that:

> [W]ind tunnel models can be used to explain as well as to predict the behavior of airplanes. To predict what the plane will do under certain conditions one observes what the model does under similar conditions. And to test competing *explanations* why the actual plane behaved as it did on some occasion, one tries to simulate the conditions and then vary them [...] (p. 115)

He then concludes:

> Thus a manipulable model, because it can be used to model counterfactual conditions, permits us to say what causes or causal factors account for the behavior of the model and thus, if we can extrapolate, permits us to say what causes or causal factors account for the behavior of whatever it is a model of. (p. 115)

One could object that merely picking out causes and causal factors is different to actually seeing the connection between those factors and "understanding *why* the cause has the effect it does" (Gordon, 1995b, p. 116). Gordon agrees that this might be a point in the example of the airplane, but wonders "just what sort of 'connection' between explanans and explanandum are we looking for when we want to understand why a person acted as she did?" (p. 116).

3.5.2 Developmental Evidence Against ST

Some criticism of simulation theory is based on developmental evidence. One of the most important study has been conducted by Perner & Howes (1992). They told children the following story: John and Mary unpack theirs bags. Since Mary has to leave, it is up to John to put the chocolate in a drawer. He has two drawers to choose from and tells Mary that he will decide later (when she is gone). After he puts the chocolate in a drawer, he goes out to play. During this time, their mother unexpectedly transfers the chocolate to the other drawer. John therefore mistakenly thinks that the chocolate is in the drawer he put it in originally.

Children's understanding is then tested with three questions. The first assesses their understanding of John's belief by asking "Where does John think the chocolate is?" The second question tests their understanding of John's self-reflection: "If we ask John: 'Do you know where the chocolate is?', what will John say?" The final question investigates Mary's reflection on John's knowledge: "If we ask Mary: 'Does John know where the chocolate is?', what will Mary say?"

The main idea behind this study is that TT and ST have different predictions about how difficult children will find it to answer these questions correctly. If children work by simulation, they would answer the questions about John like this:

> To answer the questions about John, the child has to imagine herself in John's situation, in particular, imagine herself not having seen mother transfer the chocolate to the new location. Once this hypothetical position has been taken the simulating child will find herself in a simulated false belief about where the chocolate is and in position to answer both our questions about John by assuming they were asked about herself in her simulated mental state. (Perner & Howes, 1992, p. 75–76)

Since the child is already in John's situation, both questions should be equally easy to answer. On the other hand, the question about Mary's belief will be more difficult since it requires "two levels of changes in default setting, whereas the simulation of John's mind requires but one such level of changes" (p. 76).

However, if theory theory is used instead of simulation, the second question (about John's self-reflection) will be very difficult to answer, at least according to Perner's version of TT:

> [If] children have to mentally represent John's mental state a quite different prediction follows, because explicit representation of John's subjective conviction requires formulation of a second-order state: 'John *thinks* he *knows* where the chocolate is'. This should be of comparable difficulty to representing Mary's second-order belief about John's knowledge, which we know to be substantially more difficult than representation of John's belief about the chocolate's location. (p. 76–77)

The critical question which differentiates the two theories is therefore that about how John would respond to the question about his own knowledge.

Questions	Simulation (Role-Taking)	Representing mental states
John think?	easy	easy
John say John knows?	easy	difficult
Mary say John knows?	difficult	difficult

Table taken from Perner & Howes (1992, p. 77)

The most important result of the experiment was that "there is a substantial gap between children's ability to answer the question about what John thinks and their ability to answer the self-reflection question

about what he thinks about his knowledge" (Perner & Howes, 1992, p. 79–80). The conclusion drawn from this evidence is that simulation is not used to solve the problem.

> So, the clear developmental gap between children understanding where John thinks something is and their understanding of John's insight in his belief is difficult to square with the simulation theory. (p. 82)

3.5.3 Gordon's ST and Circularity

Fuller (1995) argues that Gordon's version of simulation theory faces many circularity problems. He describes some possible circularities and offers possible responses by Gordon. The most important circularity he wants to stress "involves the last stage of simulation" (p. 25). Fuller (1995) argues that:

> It is not enough that I correctly simulate Mr Tees and go into the final stage of imaging, or pretending, that I am upset. I must also *ascribe* that state to Mr Tees. And this seems to require that I already have the, or at least *a*, concept of the mental state of being upset. (p. 24)

What Fuller seems to miss here is that according to Gordon's version of ST simulation does not require a transfer. Since you transform into the person you want to simulate, a transfer of mental states is not necessary. As the discussion develops, Fuller accepts this defense but still holds that "the problem with Gordon's account is that we are left hanging" (p. 26). Although there is no explicit reference to beliefs in Gordon's theory, Fuller objects that "there is, however, reference to appropriate simulation, and this surely requires of the ascriber that he have the *concept* of appropriate simulation and not simply the *ability* to simulate" (p. 27).

Fuller's paper lists many circularity problems which turn out not to be a problem at all. He often mentions a possible response of Gordon and the general impression is that Gordon's defense is quite well founded. However, Fuller indeed has a point to discuss possible circularity problems in Gordon's version of ST. Unfortunately, he is not very clear and does not seem to convey his points properly.

3.5.4 ST Relies on Introspection

One of the most important arguments against Harris' and Goldman's version of ST is that they depend on introspection. Gordon, for example, argues that Harris' and Goldman's theories require introspection because their simulation uses a transfer. You put *yourself* in the shoes of someone else and therefore have to perform a comparison between the other person and yourself. This, according to Gordon, needs introspective access. Gordon's theory, on the other hand, tries to solve this problem by postulating a transformation rather than a transfer.

Similarly, Carruthers (1996) claims that "they take self-knowledge of mental states for granted" (p. 28). One has to be able to recognize the beliefs, desires and intentions which are relevant for the simulation. By having access to my own mental states, I use simulation to ascribe mental states to others. According to Carruthers, this access does not happen using a theoretical description in ST. Instead, as Carruthers (1996) describes ST, 'I begin by distinguishing between one type of mental state and another purely on the basis of their intrinsic, subjectively accessible, qualities" (p. 29). This view makes ST vulnerable to "the standard objections to Cartesianism" (p. 31).

Carruthers (1996) therefore argues that Goldman and Harris "must face several difficulties" (p. 32). He gives several examples, of which I

will present one. Carruthers maintains that there are cases where we can have (and know that we have) distinct propositional episodes, and yet it is implausible that there would not be any difference in introspectible feel.

> For example, consider the difference between *intending* and *predicting* that if the party should turn out a bore then I shall fall asleep. Each state will consist, on the above account, in an image of the very same sentence – the sentence, namely, 'If the party is a bore I shall go to sleep'. So the claim be that imaging this sentence in the mode of intention is subjectively, introspectively, different from imaging it in the mode of prediction. This certainly does not fit with *my* phenomenology. Granted, I will immediately know that I have formed an intention, if I have; but not on the basis of the distinctive way that event *felt*. (Carruthers, 1996, p. 32)

This objection against ST has been sketched here very briefly since the arguments against introspection are generally known. It is important to note, however, that this criticism does not apply to Gordon's version of ST since he does not rely on introspection.

4 Conclusions

The philosophy of mind debate has gained considerable momentum during the past few years. One of the reasons for this is that the use of simulation has been proposed as an alternative to the predominant view that theory of mind is based on knowledge of a theoretical framework. Simulation theory, on the other hand, suggests that simulation – an ability – is central to mindreading and theory of mind.

Simulation theory maintains that you predict other people's behavior by putting yourself in the other person's shoes. According to Goldman, you engage in a kind of pretend play in which you feed your decision-making system with pretend inputs of beliefs and desires of the person you wish to simulate in order to predict their behavior. Then, you let your decision-making system do the rest and come to a prediction.

The significance of simulation varies between different ST theories. Gordon argues that simulation is fundamental to the mastery of psychological concepts, while other proponents of ST, such as Goldman, hold a less radical position. Goldman maintains that simulation remains the fundamental source of interpretation but acknowledges that knowledge and generalizations play an important role too.

Goldman suggests that it is plausible that evolution allowed for more than merely one strategy to develop. For him, simulation is the prime method, but the use of a theoretical framework is a possible strategy if simulation fails. This view fits in well with the evidence gained through

research about autism. Autistic people, who are generally quite bad with imagination and especially with pretend play, show an impairment of the mindreading ability. Simulation theorists, who see a connection between the reduced ability of pretend play and mindreading, suggest that autism is good evidence that we normally use simulation.

While the ability of mindreading is reduced in autistic people, there are some who can learn to master theory of mind. Following Goldman's argumentation, it has been claimed that they learn to use a theoretical framework to master easy false belief tasks. Since they still fail in harder false belief tasks and often in real world it has been suggested that this alternative strategy is not as capable as simulation.

This evidence from autism can hardly be unified with Gordon's version of ST. Since he thinks that simulation is fundamental to the master of mental concepts, there is no way that a theoretical framework can be used as an alternative to simulation. Gordon's radical view of ST, which wants to be viable without the use of any form of knowledge or theory at all, is problematic in many areas, especially when it comes to the question of how the egocentric shift is performed.

In my opinion, Goldman is on the right track striking a good balance between the ability of simulation and the use of theoretical knowledge and generalizations. We have to move into the direction of an integrated theory – a hybrid theory – which takes many different aspects into account. For example, Barresi & Moore have shown conclusively that the false belief task cannot be solved without the availability of first person information. This observation has a major impact on pure theory theories.

What we need is a theory which integrates first and third person information, and simulation and knowledge of a theory. The schema proposed by Barresi & Moore (1996) is a first step in this direction. Their

intentional schema is an intermodal perceptual and conceptual structure with the "capacity to coordinate and integrate first and third person sources of information about object-directed activities into representations that link agents to objects through intentional relations" (p. 109). Informational inputs both from oneself and other people can be integrated with the schema and the resultant representation can be applied to either yourself or another person.

Summing up, it can be said that simulation had a major impact on the philosophy of mind debate. Although simulation theories have to face severe problems and are not the panacea of philosophy of mind, their introduction has opened up new views and has shown that theory of mind is more than just the rigid use of a theoretical framework.

Bibliography

Abell, F., Happé, F., Frith, U. (2000). Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. *Journal of Cognitive Development*, 15, 1–20.

Asperger, H. (1944). Die "autistischen Psychopathen" im Kindesalter. *Archiv für Psychiatrie und Nervenkrankheiten*, 117, 76–136.

Baron-Cohen, S. (1995). *Mindblindness.* Cambridge, MA: MIT Press.

Barresi, J., Moore, C. (1995). Intentional relations and social understanding. *Behavioral and Brain Sciences*, 19, 107–154.

Blackburn, S. (1995). *Theory, observation, and drama.* In: M. Davis & T. Stone (Eds.). *Folk Psychology.* Oxford: Blackwell. 274–290.

Carruthers, P. (1996). *Simulation and self-knowledge: a defence of theory-theory.* In: P. Carruthers & P. R. Smith (Eds.). *Theories of theories of mind.* Cambridge: Cambridge University Press. 22–38.

Castelli, F., Happé, F., Frith, U., Frith, C. (2000). Movement and Mind: A Functional Imaging Study of Perception and Interpretation of Complex Intentional Movement Patterns. *NeuroImage*, 12, 314–325.

Churchland, P. M. (1990). *Eliminative Materialism and the Propositional Attitudes.* In: W. G. Lycan (Ed.). *Mind and Cognition.* Oxford: Blackwell. 206–223.

Churchland, P. M. (1991). *Folk psychology and the explanation of human behavior.* In: J. D. Greenwood (Ed.). *The future of folk psychology.* Cambridge: Cambridge University Press. 51–69.

Churchland, P. M. (1992). *Matter and Consciousness.* Cambridge, AM: MIT Press.

Currie, G. (1996). *Simulation-theory, theory-theory and the evidence from autism.* In: Carruthers, P. & Smith, P. K. (Eds.). *Theories of theories of mind.* Cambridge: Cambridge University Press. 242–256.

Davies, M., Stone, T. (1995). *Introduction.* In: M. Davies & T. Stone (Eds.). *Mental Simulation.* Oxford: Blackwell. 1–18.

Davies, M., Stone, T. (1995). *Introduction.* In: M. Davies & T. Stone (Eds.). *Folk Psychology* Oxford: Blackwell. 1–43.

Dennett, D. C. (1987). *Making sense of ourselves.* In: D. C. Dennett. *The Intentional Stance.* Cambridge, MA: MIT Press. 83–101.

Dennett, D. C. (1991). *Two contrasts: folk craft versus folk science, and belief versus opinion.* In: J. D. Greenwood (Ed.). *The future of folk psychology.* Cambridge: Cambridge University Press. 135–148.

Frith, U. (1989). *Autism: explaining the enigma.* Oxford: Basil Blackwell.

Frith, U. (1996). Cognitive explanations of autism. *Acta Pædiatr Suppl*, 416, 63–68.

Fuller, G. (1995). *Simulation and Psychological Concepts.* In: M. Davies & T. Stone (Eds.). *Mental Simulation.* Oxford: Blackwell. 19–32.

Gallese, V., Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, 2, 493–501.

Goldman, A. I. (unpublished). *Using Your Own Mind to Read Others.* Unpublished manuscript.

Goldman, A. I. (1995a). *Interpretation Psychologized.* In: M. Davis & T. Stone (Eds.). *Folk Psychology.* Oxford: Blackwell. 74–99.

Goldman, A. I. (1995b). *In Defense of the Simulation Theory.* In: M. Davis & T. Stone (Eds.). *Folk Psychology.* Oxford: Blackwell. 191–206.

Gordon, R. M. (1995). *Folk Psychology as Simulation.* In: M. Davis & T. Stone (Eds.). *Folk Psychology.* Oxford: Blackwell. 60–73.

Gordon, R. M. (1995). *The Simulation Theory: Objections and Misconceptions.* In: M. Davis & T. Stone (Eds.). *Folk Psychology.* Oxford: Blackwell. 100–122.

Gordon, R. M. (1995). *Simulation Without Introspection or Inference From Me to You.* In: M. Davies & T. Stone (Eds.). *Mental Simulation.* Oxford: Blackwell. 53–67.

Gordon, R. M. (1996). *'Radical' simulationism.* In: Carruthers, P. & Smith, P. K. (Eds.). *Theories of theories of mind.* Cambridge: Cambridge University Press. 11–21.

Greenwood, J. D. (1991). *Introduction: Folk psychology and scientific psychology.* In: J. D. Greenwood (Ed.). *The future of folk psychology.* Cambridge: Cambridge University Press. 1–21.

Happé, F. (1994). *Autism: an introduction to psychological theory.* London: UCL Press.

Happé, F. (1999). Autism: cognitive deficit or cognitive style? *Trends in Cognitive Sciences*, 3 (6), 216–222.

Harris, P. L. (1995). *From Simulation to Folk Psychology: The Case for Development.* In: M. Davis & T. Stone (Eds.). *Folk Psychology.* Oxford: Blackwell. 207–231.

Heal, J. (1995). *How to Think About Thinking.* In: M. Davies & T. Stone (Eds.). *Mental Simulation.* Oxford: Blackwell. 33–52.

Humphrey, N. (1984). *Consciousness regained.* Oxford: Oxford University Press.

Kahneman, D., Tversky, A. (1982). *The simulation heuristic.* In: D. Kahneman, P. Slovic & A. Tversky (Eds.). *Judgment Under Uncertainty.* Cambridge: Cambridge University Press. 201–208.

Kanner, L. (1943). Autistic disturbances of affective contact. *Nervous Child*, 2, 217–250.

Leslie, A. M., German, T. P. (1995). *Knowledge and Ability in "Theory of Mind": One-eyed Overview of a Debate.* In: M. Davies & T. Stone (Eds.). *Mental Simulation.* Oxford: Blackwell. 123–150.

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

Perner, J., Howes, D. (1992). 'He Thinks He Knows': And More Developmental Evidence Against the Simulation (Role Taking) Theory. *Mind & Language*, 7, 72–86.

Premack, D., Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 4, 515–526.

Rizzolatti, G., Arbib, M. A. (1998). Language within our grasp. *Trends in Neuroscience*, 21, 188–194.

Shepard, R. N., Metzler, J. (1971). Mental Rotation of Three-Dimensional Objects. *Science*, 171. 701–703.

Stich, S., Nichols, S. (1995). *Second Thoughts on Simulation.* In: M. Davies & T. Stone (Eds.). *Mental Simulation.* Oxford: Blackwell. 87–108.

Wimmer, H., Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13, 103–128.

Yue, G., Cole, K. (1992). Strength increases from the motor program: Comparison of training with maximal voluntary and imagined muscle contractions. *Journal of Neurophysiology*, 67. 1114–1123.

`$Id: text.tex,v 1.113.2.5 2002/03/03 22:09:08 tbm Exp $`