Mind in practice : a pragmatic and interdisciplinary account of intersubjectivity

Bruin, L.C. de

2.

Simulation Theory

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

- Doyle 1986

Folk psychology is simulation

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

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

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

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

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

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

2.1 Making sense of simulation

Simulation theory according to Goldman

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

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

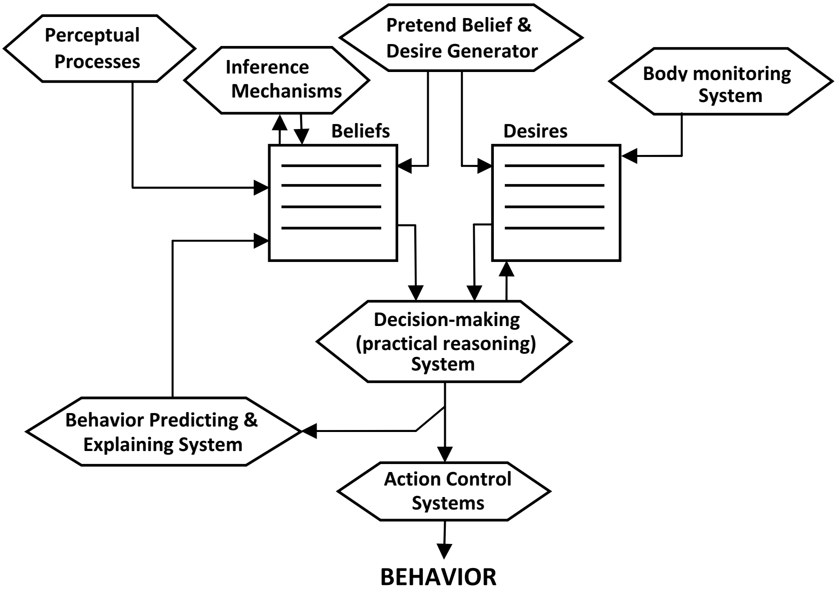
Perceptual Processes
Inference Mechanisms
Pretend Belief & Desire Generator
Body monitoring System
Beliefs
Desires
Decision-making (practical reasoning) System
Behavior Predicting & Explaining System
Action Control Systems
BEHAVIOR

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

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

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

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

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

Some initial complications

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

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

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

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

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

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

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

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

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

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

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

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

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

The argument from phenomenology revisited

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

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

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

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

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

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

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

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

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

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

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

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

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

Simulation theory according to Gordon

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

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

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

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

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

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

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

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

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

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

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

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

Simulation theory according to Heal

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

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

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

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

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

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

A threat of collapse and the return of folk psychological principles

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.2 Assessing the empirical evidence

Again, the false belief test

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

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

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

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

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

Imitation and pretend play

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

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

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

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

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

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

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

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

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

Tacit simulation: how low can we go?

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

Initially, Gallese and Goldman (1989) conjectured that such a mirror neuron system could be seen as a 'primitive version, or possibly a precursor in phylogeny, of a simulation heuristic that might underlie mindreading' (p. 498). Mirror neuron activity seemed to be 'nature's way of getting the observer into the same "mental shoes" as the target - exactly what the conjectured simulation heuristic aims to do' (ibid.). The mirror neuron system supported at least a kind of low-level simulation, so it was thought, but it probably also paved the way for high-level simulation in all its glory.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.3 Simulation, anyone?

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

Simulation summarized

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

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

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

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

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