Mind in practice : a pragmatic and interdisciplinary account of intersubjectivity

Bruin, L.C. de

1.

Theory Theory

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

- Sellars 1963

Mindreading

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

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

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

of the term 'it'. On the contrary, he immediately realizes that the word ‘it’ refers to the gun in his hand. But how is he able to do this?

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

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

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

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

unable to make sense of others. The attribution of mental states to others is our natural way of understanding the social environment. In the words of Sperber (1993), 'attribution of mental states is to humans as echolocation is to the bat'. Without mindreading, the other mind remains a mystery.

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

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

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

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

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

1.1 Folk psychology as theory

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

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

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

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

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

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

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

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

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

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

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

1.2 A taste of TT

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

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

prediction of another organism's behavior (cf. Baron-Cohen 1995). This view is often complemented by the 'Machiavellian intelligence' hypothesis, according to which a primary selection pressure driving human brain development was strategic interaction, with social competition leading to increasingly sophisticated mindreading mechanisms (e.g. Byrne and Whiten 1988).

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

Innateness and the problem of learning

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

account for the abstract, complex nature of knowledge, but cannot explain, and therefore denies, the fact that we learn. The latter is able to explain learning, but can't explain, and so denies, the fact that our knowledge is so far removed from experience.

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

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

the boat to keep it seaworthy. In the end, however, we may end up with not a single plank or rivet from the original structure, and the process may go on indefinitely' (2003, p.242).

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

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

Belief-desire psychology and the problem of context-sensitivity

Although it is often suggested that folk psychology includes much more than the ability to make sense of others in terms of beliefs and desires, there is a strong consensus that it

should at the very least include this ability.[6] Since philosophical orthodoxy has it that individual beliefs cannot cause actions on their own, and lone desires are aimless without guiding beliefs, it is thought that we need to discover a proper *combination* of them in order to understand others and predict or explain their actions.

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

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

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

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

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

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

Folk psychological principles 'ain't in the head'

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

Some philosophers have argued that these generalizations might be usefully thought of as a term-introducing theory which implicitly defines terms such as 'believe', 'want' and 'desire' (e.g., Lewis 1972). Braddon-Mitchell and Jackson (2007), for example, follow this line and argue that the existence of folk psychological rules 'does not, of course, mean that we must have a theory […] explicitly worked out in our minds, but somehow hidden from

view and guiding our actions from its hiding place. Rather, it means that our responses to situations and our [folk psychological] judgments […] are governed in most cases by our existing networks of interrelated powers of discrimination' (p.63).

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

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

Moreover, the appeal to social practice can also be used to mount an extra argument *against* both externalist and internalist versions of TT. For example, in their evaluation of the myth of Jones and its significance for TT, Stich and Ravenscroft (1996) point out that, as Sellars tells the story, Jones self-consciously develops a folk psychological theory and explicitly teaches it to his compatriots. But Stich and Ravenscroft observe that nothing like that seems to go on in our current social practice: 'We don't explicitly teach our children a theory that enables them to apply mental terms to other people. Indeed, unlike Jones and

his friends, we are not even able to state the theory, let alone teach it. If you ask your neighbor to set out the principles of the theory of the mind that she has taught her children, she won't have the foggiest idea what you're talking about' (pp.121-2). A similar argument against TT is made by Goldman (1989), who also wonders how children might get a grip on a theory as complex and sophisticated as the one that TT attributes to them: 'One possible mode of acquisition is cultural transmission (e.g. being taught them explicitly by their elders). This is clearly out of the question, though, since only philosophers have ever tried to articulate the laws, and most children have no exposure to philosophers' (pp.167-8). This brings us to a broader, more encompassing phenomenological argument against TT.

1.3 Where is the theory in TT?

The argument from phenomenology

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

interactions with others break down and we have trouble understanding them. However, these cases are the exception rather than the rule. Normally, intersubjectivity does not involve a 'detached or abstract observational stance', since our understanding of others 'is poorly described as involving the formulation of a theoretical hypothesis' (ibid.).

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

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

The appeal to tacit theory

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

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

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

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

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

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

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

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

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

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

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

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

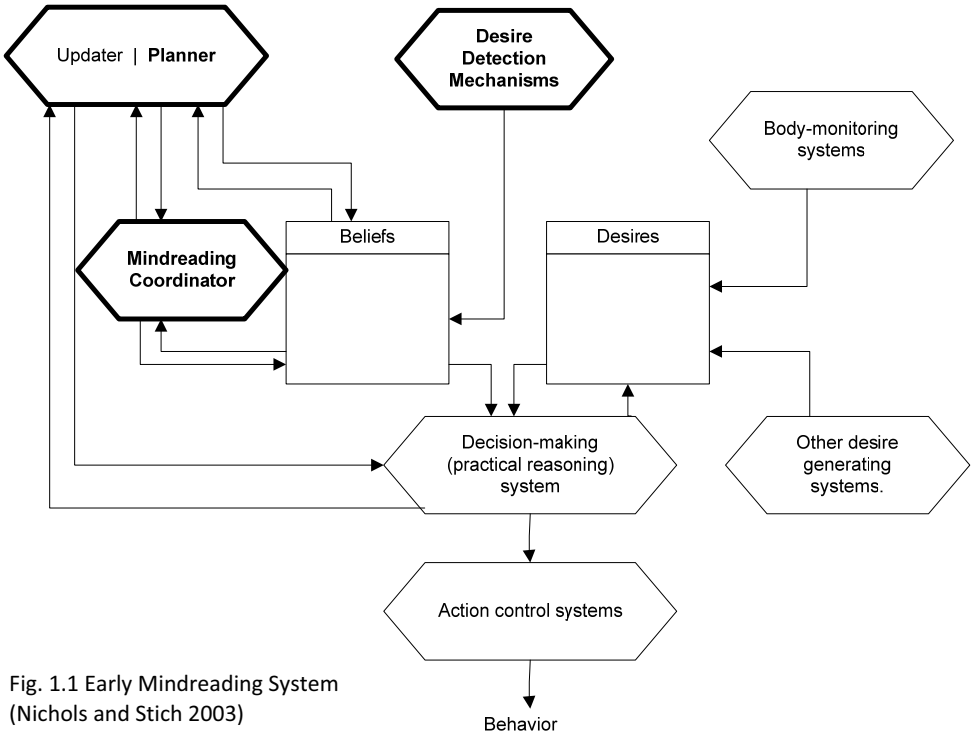

Fig. 1.1 Early Mindreading System (Nichols and Stich 2003)

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

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

1.4 Scientific evidence for TT

Theorizing chimps

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

Premack and Woodruff's suggestion has been very influential, and their article set the stage for a major episode in research on theory of mind in children. As Gopnik (1993) observes, 'in the last few years there has been an explosion of interest in children's ideas about the mind' (p.3). In a similar fashion (but on a more critical note), Reddy and Morris (2004) remark that it 'is difficult to write today about understanding people without reference to the words "theory of mind". An incredible 1 percent of academic publications in psychology in 2003-4 that refer to infants or children also refer to the term "theory of

mind". And the manner in which the term is used is awesomely matter-of-fact-with a taken-for-grantedness hitherto reserved for those other staples of psychology such as "growth spurt", "toilet training", "short-term memory" and "secure attachment"' (p.647).

The false belief test

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

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

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

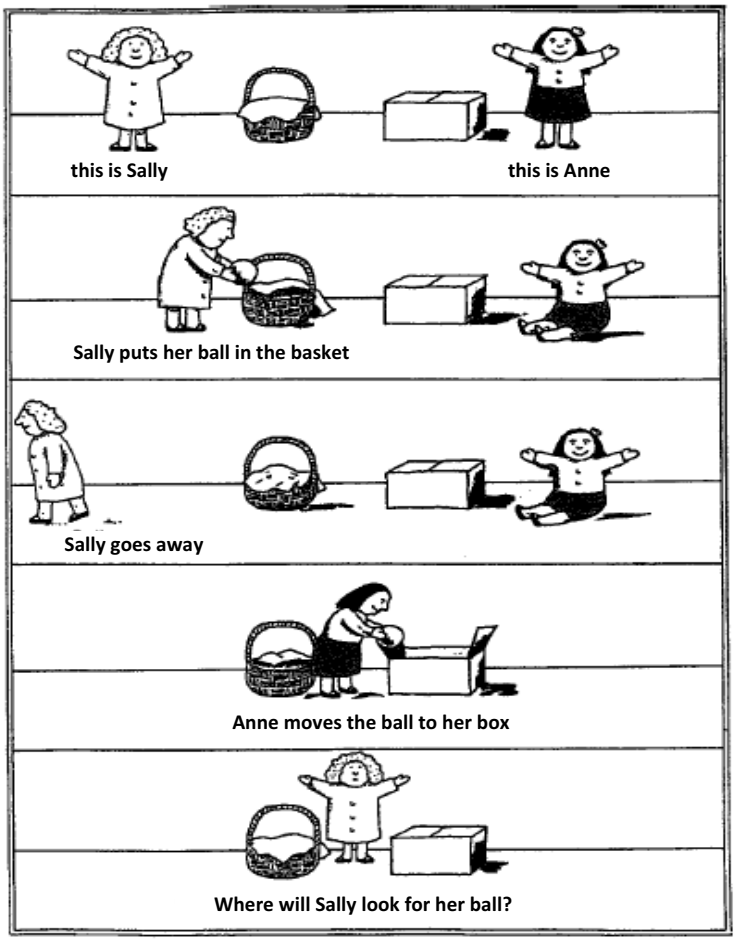

Fig. 1.2 The Sally-Anne False belief test

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

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

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

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

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

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

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

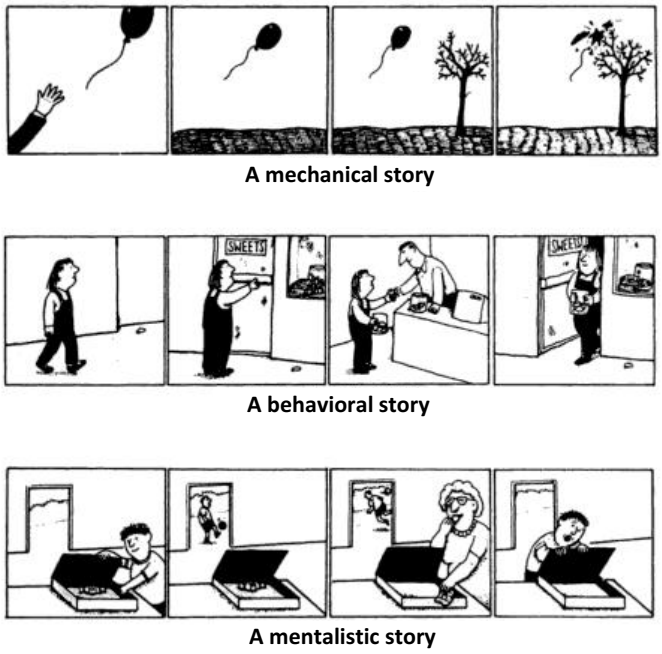

Fig. 1.3 Three types of picture sequences

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

Neuroscientists have tried to trace the neurobiological roots of this impediment. Castelli et al. (2002), for example, PET-scanned autistic and normal subjects while they were watching animated sequences. The animations depicted two triangles moving about on a screen in three different conditions: moving randomly, moving in a goal-directed

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

A question of interpretation

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

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

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

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

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

ignorance, their emotions and desires, and demonstrate sufficient understanding that other people's likes and desires may be different from the child's own (cf. Dunn 1988).

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

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

Gopnik and Meltzoff (1994) argue that most of the developmental evidence indeed points in this direction. They claim that the evidence suggests that there is an extensive parallelism between children's understanding of their own mental states and their understanding of the mental states of others: 'In each of our studies, children's reports of their own immediately past psychological states are consistent with their accounts of the psychological states of others. When they can report and understand the psychological states of others, in the cases of pretense, perception and imagination, they report having had those psychological states themselves. When they cannot report and understand the psychological states of others, in the case of beliefs and source, they do not report that they had those states themselves' (pp.179-80). I already mentioned the 'Smarties' test, in which children are presented with a candy box full of pencils. Gopnik and Astington (1988) have shown that 3-year-old children not only predict that others will think there are pencils in this candy box (without having looked into it), but also that they make the same error

when they are asked about their *own* immediately past false beliefs. In this case, they report that they *already thought* that there were pencils in the box. According to Gopnik (1993), this proves that there is no such thing as a privileged access to our own mind.[13]

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

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

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

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

Belief-desire psychology in low-level action understanding?

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

Explicit versions of the false belief test clearly show that something new and important happens at the age of 4 years, and that this something is somewhat consistent with certain assumptions of TT. However, since these tests are designed to capture a very specialized mode of intersubjectivity, they cannot be used to validate a TT approach to intersubjectivity *in general.* This is not because the evidence is lacking, but it rather has to do with the proper interpretation of this evidence and the questioning of a certain picture of

intersubjectivity that is presupposed by TT. In implicit false belief tests of the Onishi and Baillargeon kind, it also concerns the appropriate *level of explanation*. Suppose we grant TT that our understanding of others is facilitated by a *tacit* folk psychological theory. The question still remains whether it is possible to map belief-desire processes directly onto the sub-personal level, using personal level vocabulary as if nothing has changed.[14] At the very least, we should carefully explain what we mean by 'tacit' beliefs and/or desires.

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

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

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

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

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

propositional attitude psychology. Knowledge of how these propositional attitudes interrelate with one another 'comprises what we might think of as the "core principles" of intentional psychology' (p.29).

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

1.5 A number of pressing TT problems

A short summary

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

vocabulary. However, attempts to articulate these putative laws or 'platitudes' have been notably weak, and externalist TT also lacks a developmental story.

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

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

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

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

Folk psychology is false as a theory...

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

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

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

According to Churchland, these generalizations are fallible empirical hypotheses. The mental concepts they employ are defined by their place in the overall system of laws. They

are the theoretical terms of a theoretical framework, and their meanings are fixed by the set of generalizations in which they figure. 'Theoretical terms do not, in general, get their meanings from single, explicit definitions stating conditions necessary and sufficient for application. They are implicitly defined by the network of principles that embed them' (1988, p.56). Theoretical terms primarily have a predictive and explanatory function, and Churchland argues that this is also their main value.

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

...but folk psychology might not be a theory

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

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

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

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

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