Forthcoming in *New Essays in Philosophy of Language and Mind*, a supplemental volume of the *Canadian Journal of Philosophy*, eds. M. Ezcurdia, R. Stainton & C. Viger.

Reading One's Own Mind: Self-Awareness and Developmental Psychology

Shaun Nichols
Stephen Stich

1. Introduction

The idea that we have special access to our own mental states has a distinguished philosophical history. Philosophers as different as Descartes and Locke agreed that we know our own minds in a way that is quite different from the way in which we know other minds. In the latter half of the 20th century, however, this idea came under serious attack, first from philosophy (Sellars 1956) and more recently from developmental psychology.[1] The attack from developmental psychology arises from the growing body of work on "mindreading", the process of attributing mental states to people (and other organisms). During the last 15 years, the processes underlying mindreading have been a major focus of attention in cognitive and developmental psychology. Most of this work has been concerned with the processes underlying the attribution of mental states to *other* people. However, a number of psychologists and philosophers have also proposed accounts of the mechanisms underlying the attribution of mental states to *oneself*. This process of *reading one's own mind* or *becoming self-aware* will be our primary concern in this paper.

We'll start by examining what is probably the most widely held account of self-awareness in this literature, the "Theory Theory" (TT). The basic idea of the TT of self-awareness is that one's access to one's own mind depends on the same cluster of cognitive mechanisms that plays a central role in attributing mental states to others. Those mechanisms include a body of information about psychology, a Theory of Mind (ToM). Though many authors have endorsed the Theory Theory of self-awareness (Gopnik 1993, Gopnik & Wellman 1994, Gopnik & Meltzoff 1994, Perner 1991, Wimmer & Hartl 1991, Carruthers 1996, C.D. Frith 1994, U. Frith & Happé 1999), it is our contention that advocates of this account of self-awareness have left their theory seriously under-described. In the next section, we'll suggest three different ways in which the TT account might be elaborated, all of which have significant shortcomings. In section 3, we'll present our own theory of self-awareness, the Monitoring Mechanism Theory, and compare its merits to those of the TT. Theory Theorists argue that the TT is supported by evidence about psychological development and psychopathologies. In section 4 we will review the developmental arguments and try to show that none of the evidence favors the TT over our Monitoring Mechanism Theory. Indeed, we'll maintain

[1] For more on Sellars' role in this challenge to the traditional view, see Stich & Ravenscroft (1994).

that a closer look at the evidence on development actually provides arguments *against* the TT.[2] Elsewhere, we consider the evidence from psychopathologies (Nichols & Stich 2002, forthcoming). Nichols & Stich (2002) is intended as a companion piece to this article. There too, we argue that despite the advertisements, the evidence – in that paper our focus is on the evidence concerning psychopathologies – poses a problem for the Theory Theory. We should note that there is considerable overlap between the present paper and Nichols & Stich (2002). In both papers, we consider whether the evidence favors the Theory Theory or the Monitoring Mechanism theory, and the theoretical background against which the arguments are developed is largely the same in both papers. So the (cherished) reader familiar with the companion paper (Nichols & Stich 2002) might skip to section 4, where we take on the developmental arguments. We now turn to the task of setting out the background for the debate.

Mindreading skills, in both the first person and the third person cases, can be divided into two categories which, for want of better labels, we'll call *detecting* and *reasoning*.

a. *Detecting* is the capacity to *attribute* current mental states to someone.

b. *Reasoning* is the capacity to *use* information about a person's mental states (typically along with other information) to make predictions about the person's past and future mental states, her behavior, and her environment.

So, for instance, one might *detect* that another person wants ice cream and that the person thinks the closest place to get ice cream is at the corner shop. Then one might *reason* from this information that, since the person wants ice cream and thinks that she can get it at the corner shop, she will go to the shop. The distinction between detecting and reasoning is an important one because some of the theories we'll be considering offer integrated accounts on which detecting and reasoning are explained by the same cognitive mechanism. Other theories, including ours, maintain that in the first person case, these two aspects of mindreading are subserved by different mechanisms.

Like the other authors we'll be considering, we take it to be a requirement on theories of self-awareness that they offer an explanation for:

i) the obvious facts about self-attribution (e.g. that normal adults do it easily and often, that they are generally accurate, and that they have no clear idea of how they do it)

---

[2] Although the Theory Theory is the most prominent account of self-awareness in this literature, there are two other widely discussed theories of self-awareness to be found in the recent literature: Alvin Goldman's (1993a, 1993b, 1997, 2000) phenomenological account and Robert Gordon's "ascent routine" account (Gordon 1995, 1996). We think that neither of these accounts can capture the basic facts about self-awareness, and we make our case against them in Nichols & Stich (2002).

ii) the often rather un-obvious facts about self-attribution that have been uncovered by cognitive and developmental psychologists (e.g., Gopnik & Slaughter 1991, Ericsson & Simon 1993, Nisbett & Wilson 1977).

However, we *do not* take it to be a requirement on theory building in this area that the theory address philosophical puzzles that have been raised about knowledge of one's own mental states. In recent years, philosophers have had a great deal to say about the link between content externalism and the possibility that people can have privileged knowledge about their own propositional attitudes (e.g., McLaughlin & Tye 1998)[3]. These issues are largely orthogonal to the sorts of questions about underlying mechanisms that we will be discussing in this paper, and we have nothing at all to contribute to the resolution of the philosophical puzzles posed by externalism. But in the unlikely event that philosophers who worry about such matters agree on solutions to these puzzles, we expect that the solutions will fit comfortably with our theory.

There is one last bit of background that needs to be made explicit before we begin. The theory we'll set out will help itself to two basic assumptions about the mind. We call the first of these *the basic architecture assumption.* What it claims is that a well known commonsense account of the architecture of the cognitive mind is largely correct, though obviously incomplete. This account of cognitive architecture, which has been widely adopted both in cognitive science and in philosophy, maintains that in normal humans, and probably in other organisms as well, the mind contains two quite different kinds of representational states, beliefs and desires. These two kinds of states differ "functionally" because they are caused in different ways and have different patterns of interaction with other components of the mind. Some beliefs are caused fairly directly by perception; others are derived from pre-existing beliefs via processes of deductive and non-deductive inference. Some desires (like the desire to get something to drink or the desire to get something to eat) are caused by systems that monitor various bodily states. Other desires, sometimes called "instrumental desires" or "sub-goals," are generated by a process of practical reasoning that has access to beliefs and to pre-existing desires. In addition to generating sub-goals, the practical reasoning system must also determine which structure of goals and sub-goals is to be acted upon at any time. Once made, that decision is passed on to various action controlling systems whose job it is to sequence and coordinate the behaviors necessary to carry out the decision. Figure 1 is a sketch of the basic architecture assumption.

FIGURE 1 ABOUT HERE

---

[3]Content externalism is the view that the content of one's mental states (what the mental states are about) is determined at least in part by factors external to one's mind. In contemporary analytic philosophy, the view was motivated largely by Putnam's Twin Earth thought experiments (Putnam 1975) that seem to show that two molecule for molecule twins can have thoughts with different meanings, apparently because of their different external environments.

We find diagrams like this to be very helpful in comparing and clarifying theories about mental mechanisms, and we'll make frequent use of them in this paper. It is important, however, that the diagrams not be misinterpreted. Positing a "box" in which a certain category of mental states are located is simply a way of depicting the fact that those states share an important cluster of causal properties that are not shared by other types of states in the system. There is no suggestion that all the states in the box share a spatial location in the brain. Nor does it follow that there can't be significant and systematic differences among the states within a box. When it becomes important to emphasize such differences, we use boxes within boxes or other obvious notational devices. All of this applies as well to processing mechanisms, like the inference mechanism and the practical reasoning mechanism, which we distinguish by using hexagonal boxes.

Our second assumption, which we'll call *the representational account of cognition*, maintains that beliefs, desires and other propositional attitudes are relational states. To have a belief or a desire with a particular content is to have a representation token with that content stored in the functionally appropriate way in the mind. So, for example, to believe that Socrates was an Athenian is to have a representation token whose content is *Socrates was an Athenian* stored in one's Belief Box, and to desire that it will be sunny tomorrow is to have a representation whose content is *It will be sunny tomorrow* stored in one's Desire Box. Many advocates of the representational account of cognition also assume that the representation tokens subserving propositional attitudes are linguistic or quasi-linguistic in form. This additional assumption is no part of our theory, however. If it turns out that some propositional attitudes are subserved by representation tokens that are not plausibly viewed as having a quasi-linguistic structure, that's fine with us.

We don't propose to mount any defense of these assumptions here. However, we think it is extremely plausible to suppose that the assumptions are shared by most or all of the authors whose views we will be discussing.

2. The Theory Theory

As noted earlier, the prevailing account of self-awareness is the Theory Theory. Of course, the prevailing account of how we understand *other minds* is also a Theory Theory. Before setting out the Theory Theory account of reading one's own mind, it's important to be clear about how the Theory Theory proposes to explain our capacity to read other minds.[4]

[4]In previous publications on the debate between the Theory Theory and Simulation Theory, we have defended the Theory Theory of how we understand other minds (Stich & Nichols 1992; Stich & Nichols 1995; Nichols et al. 1995; Nichols et al 1996). More recently, we've argued that the Simulation/Theory Theory debate has outlived its usefulness, and productive debate will require more detailed proposals and sharper

2.1. The Theory Theory account of reading other people's minds

According to the Theory Theory, the capacity to *detect* other people's mental states relies on a theory-mediated inference. The theory that is invoked is a Theory of Mind which some authors (e.g. Fodor 1992; Leslie 1994b) conceive of as a special purpose body of knowledge housed in a mental module, and others (e.g. Gopnik & Wellman 1994) conceive of as a body of knowledge that is entirely parallel to other theories, both common sense and scientific. For some purposes the distinction between the modular and the just-like-other-(scientific)-theories versions of the Theory Theory is of great importance. But for our purposes it is not. So in most of what follows we propose to ignore it (but see Stich & Nichols 1998). On all versions of the Theory Theory, when we detect another person's mental state, the theory-mediated inference can draw on perceptually available information about the behavior of the target and about her environment. It can also draw on information stored in memory about the target and her environment. A sketch of the mental mechanisms invoked in this account is given in Figure 2.

FIGURE 2 ABOUT HERE

The theory that underlies the capacity to *detect* other people's mental states also underlies the capacity to *reason* about other people's mental states and thereby predict their behavior. Reasoning about other people's mental states is thus a theory-mediated inference process, and the inferences draw on beliefs about (*inter alia*) the target's mental states. Of course, some of these beliefs will themselves have been produced by detection inferences. When detecting and reasoning are depicted together we get Figure 3.

FIGURE 3 ABOUT HERE

2.2. Reading one's own mind: Three versions of the TT account.

The Theory Theory account of how we read other minds can be extended to provide an account of how we read our own minds. Indeed, both the Theory Theory for understanding other minds and the Theory Theory for self-awareness seem to have been first proposed in the same article by Wilfrid Sellars (1956). The core idea of the TT account of self-awareness is that the process of reading one's own mind is largely or entirely parallel to the process of reading someone else's mind. Advocates of the Theory Theory of self-awareness maintain that knowledge of one's own mind, like knowledge of other minds, comes from a theory-mediated inference, and the theory that mediates the

---

distinctions (Stich & Nichols 1997; Nichols & Stich 1998). In this paper we've tried to sidestep these issues by granting the Theory Theorist as much as possible. We maintain that even if *all* attribution and reasoning about other minds depends on theory, that still won't provide the Theory Theorist with the resources to accommodate the facts about self-awareness.

inference is the same for self and other – it's the Theory of Mind. In recent years many authors have endorsed this idea; here are two examples:

> Even though we seem to perceive our own mental states directly, this direct perception is an illusion. In fact, our knowledge of ourselves, like our knowledge of others, is the result of a theory, and depends as much on our experience of others as on our experience of ourselves (Gopnik & Meltzoff 1994, 168).

> …if the mechanism which underlies the computation of mental states is dysfunctional, then self-knowledge is likely to be impaired just as is the knowledge of other minds. The logical extension of the ToM [Theory of Mind] deficit account of autism is that individuals with autism may know as little about their own minds as about the minds of other people. This is not to say that these individuals lack mental states, but that in an important sense they are unable to reflect on their mental states. Simply put, they lack the cognitive machinery to represent their thoughts and feelings as thoughts and feelings (Frith & Happé 1999, 7).

As we noted earlier, advocates of the TT account of self-awareness are much less explicit than one would like, and unpacking the view in different ways leads to significantly different versions of the TT account. But all of them share the claim that the processes of reasoning about and detecting one's own mental states will parallel the processes of reasoning about and detecting others' mental states. Since the process of *detecting* one's own mental states will be our focus, it's especially important to be very explicit about the account of detection suggested by the Theory Theory of self-awareness. According to the TT:

> i. Detecting one's own mental states is a theory-mediated inferential process. The theory, here as in the third person case, is ToM (either a modular version or a just-like-other-(scientific)-theories version or something in between).
>
> ii. As in the 3rd person case, the capacity to detect one's own mental states relies on a theory-mediated inference which draws on perceptually available information about one's own behavior and environment. The inference also draws on information stored in memory about oneself and one's environment.

At this point the TT account of self-awareness can be developed in at least three different ways. So far as we know, advocates of the TT have never taken explicit note of the distinction. Thus it is difficult to determine which version a given theorist would endorse.

2.2.1. Theory Theory Version 1

Theory Theory version 1 (for which our code name is *the crazy version*) proposes to maintain the parallel between detecting one's own mental states and detecting another person's mental states quite strictly. The *only* information used as evidence for the inference involved in detecting one's own mental state is the information provided by perception (in this case, perception of oneself) and by one's background beliefs (in this case, background beliefs about one's own environment and previously acquired beliefs about one's own mental states). This version of TT is sketched in Figure 4.

FIGURE 4 ABOUT HERE

Of course, we typically have much more information about our own minds than we do about other minds, so even on this version of the Theory Theory we may well have a *better* grasp of our own mind than we do of other minds (see e.g., Gopnik 1993, 94). However, the mechanisms underlying self-awareness are supposed to be the same mechanisms that underlie awareness of the mental states of others. Thus this version of the TT denies the widely held view that an individual has some kind of special or privileged access to his own mental states.

We are reluctant to claim that anyone actually advocates this version of the TT, since we think it is a view that is hard to take seriously. Indeed, the claim that *perception of one's own behavior* is the prime source of information on which to base inferences about one's own mental states reminds us of the old joke about the two behaviorists who meet on the street. One says to the other, "You're fine. How am I?" The reason the joke works is that it seems patently absurd to think that perception of one's behavior is the best way to find out how one is feeling. It seems obvious that people can sit quietly without exhibiting any relevant behavior and report on their current thoughts. For instance, people can answer questions about current mental states like "what are you thinking about?". Similarly, after silently working a problem in their heads, people can answer subsequent questions like "how did you figure that out?". And we typically assume that people are correct when they tell us what they were thinking or how they just solved a problem. Of course, it's not just one's current and immediately past *thoughts* that one can report. One can also report one's own current desires, intentions, and imaginings. It seems that people can easily and reliably answer questions like: "what do you want to do?"; "what are you going to do?"; "what are you imagining?" People who aren't exhibiting much behavior at all are often able to provide richly detailed answers to these questions. These more or less intuitive claims are backed by considerable empirical evidence from research programs in psychology (see, e.g., Ericsson & Simon 1993).

So, both commonsense and experimental studies confirm that people can sit quietly, exhibiting next to no overt behavior, and give detailed, accurate self-reports about their mental states. In light of this, it strikes us as simply preposterous to suggest that the reports people make about their own mental states are being inferred from perceptions of their own behavior and information stored in memory. For it's simply absurd to suppose that there is enough behavioral evidence or information stored in memory to serve as a basis for accurately answering questions like "what are you thinking about now?" or "how did you solve that math problem?". Our ability to answer questions like these indicates that Version 1 of the Theory Theory of self-awareness can't be correct since it can't accommodate some central cases of self-awareness.

2.2.2. Theory Theory Version 2

Version 2 of the Theory Theory (for which our code name is *the under-described version*) allows that in using ToM to infer to conclusions about one's own mind there is information available *in addition to* the information provided by perception and one's background beliefs. This additional information is available only in the 1st person case, not in the 3rd person case. Unfortunately, advocates of the TT say very little about what this alternative source of information is. And what little they do say about it is unhelpful to put it mildly. Here, for instance, is an example of the sort of thing that Gopnik has said about this additional source of information:

> One possible source of evidence for the child's theory may be first-person psychological 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"* (Gopnik 1993, 11).

We have no serious idea what the "Cartesian buzz" is, or how one would detect it. Nor do we understand how detecting the Cartesian buzz will enable the ToM to infer to conclusions like: *I want to spend next Christmas in Paris* or *I believe that the Brooklyn Bridge is about eight blocks south of the Manhattan Bridge*. Figure 5 is our attempt to sketch Version 2 of the TT account.

FIGURE 5 ABOUT HERE

We won't bother to mount a critique against this version of the account, apart from observing that without some less mysterious statement of what the additional source(s) of information are, the theory is too incomplete to evaluate.

2.2.3. Theory Theory Version 3

There is, of course, one very natural way to spell out what's missing in Version 2. What is needed is some source of information that would help a person form beliefs

(typically true beliefs) about his own mental states. The obvious source of information would be the mental states themselves. So, on this version of the TT, the ToM has access to information provided by perception, information provided by background beliefs, *and information about the representations contained in the Belief Box, the Desire Box, etc.* This version of the TT is sketched in Figure 6.

FIGURE 6 ABOUT HERE

Now at this juncture one might wonder why the ToM is *needed* in this story. If the mechanism subserving self-awareness has access to information about the representations in the various attitude boxes, then ToM has no serious work to do. So why suppose that it is involved at all? That's a good question, we think. And it's also a good launching pad for our theory. Because on our account Figure 6 has it wrong. In detecting one's own mental states, the flow of information is *not* routed through the ToM. Rather, the process is subserved by a separate self-monitoring mechanism.

3. Reading one's own mind: The Monitoring Mechanism Theory

In constructing our theory about the process that subserves self-awareness we've tried to be, to borrow a phrase from Nelson Goodman, (1983, 60) "refreshingly non-cosmic". What we propose is that we need to add another component or cluster of components to the basic picture of cognitive architecture, a mechanism (or mechanisms) that serves the function of monitoring one's own mental states.

3.1. The Monitoring Mechanism and propositional attitudes

Recall what the theory of self-awareness needs to explain. The basic facts are that when normal adults believe that p, they can quickly and accurately form the belief *I believe that p*; when normal adults desire that p, they can quickly and accurately form the belief *I desire that p*; and so on for the rest of the propositional attitudes. In order to implement this ability, no sophisticated Theory of Mind is required. All that is required is that there be a Monitoring Mechanism (MM) (or perhaps a set of mechanisms) 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 would work in much the same way to produce

representations of one's own desires, intentions, and imaginings.[5] This account of the process of self-awareness is sketched in Figure 7.

FIGURE 7 ABOUT HERE

Although we propose that the MM is a special mechanism for detecting one's own mental states, we maintain that there is no special mechanism for what we earlier called *reasoning about* one's own mental states. Rather, reasoning about one's own mental states depends on the same Theory of Mind as reasoning about others' mental states. As a result, our theory (as well as the TT) predicts that, *ceteris paribus*, where the ToM is deficient or the relevant information is unavailable, subjects will make mistakes in reasoning about their own mental states as well as others. This allows our theory to accommodate findings like those presented by Nisbett & Wilson (1977). They report a number of studies in which subjects make mistakes about their own mental states. However, the kinds of mistakes that are made in those experiments are typically not mistakes in *detecting* one's own mental states. Rather, the studies show that subjects make mistakes in *reasoning about* their own mental states. The central findings are that subjects sometimes attribute their behavior to inefficacious beliefs and that subjects sometimes deny the efficacy of beliefs that are, in fact, efficacious. For instance, Nisbett & Schacter (1966) found that subjects were willing to tolerate more intense shocks if the subjects were given a drug (actually a placebo) and told that the drug would produce heart palpitations, irregular breathing and butterflies in the stomach. Although being told about the drug had a significant effect on the subjects' willingness to take shocks, most subjects denied this. Nisbett & Wilson's explanation of these findings is, plausibly enough, that subjects have an incomplete theory regarding the mind and that the subjects' mistakes reflect the inadequacies of their theory (Nisbett & Wilson 1977). This explanation of the findings fits well with our account too. For on our account, when trying to figure out the *causes* of one's own behavior, one must reason about mental states, and this process is mediated by the ToM. As a result, if the ToM is not up to the task, then people will make mistakes in reasoning about their own mental states as well as others' mental states.

In this paper, we propose to remain agnostic about the extent to which ToM is innate. However, we do propose that the MM (or cluster of MMs) is innate and comes on line fairly early in development – significantly before ToM is fully in place. During the period when the Monitoring Mechanism is up and running but ToM is not, the representations that the MM produces can't do much. In particular, they can't serve as premises for reasoning about mental states, since reasoning about mental states is a process mediated by ToM. So, for example, ToM provides the additional premises (or

[5]Apart from the cognitive science trappings, the idea of an internal monitor goes back at least to David Armstrong (1968) and has been elaborated by William Lycan (1987) among others. However, much of this literature has become intertwined with the attempt to determine the proper account of consciousness, and that is not our concern at all. Rather, on our account, the monitor is just a rather simple information-processing mechanism that generates explicit representations about the representations in various components of the mind and inserts these new representations in the Belief Box.

the special purpose inferential strategies) that enable the mind to go from premises like *I want q* to conclusions like: *If I believed that doing A was the best way to get q, then (probably) I would want to do A.* Thus our theory predicts that young children can't reason about their own beliefs in this way.

Although we want to leave open the extent to which ToM is innate, we maintain (along with many Theory Theorists) that ToM comes on line only gradually. As it comes on line, it enables a richer and richer set of inferences from the representations of the form *I believe (or desire) that p* that are produced by the MM. Some might argue that early on in development, these representations of the form *I believe that p* don't really count as having the content: *I believe that p*, since the concept (or "proto-concept") of belief is too inferentially impoverished. On this view, it is only after a rich set of inferences becomes available that the child's *I believe that p* representations really count as having the content: *I believe that p*. To make a persuasive case for or against this view, one would need a well motivated and carefully defended theory of content for concepts. And we don't happen to have one. (Indeed, at least one of us is inclined to suspect that much recent work aimed at constructing theories of content is deeply misguided [Stich 1992, 1996].) But, with this caveat, we don't have any objection to the claim that early *I believe that p* representations don't have the content: *I believe that p*. If that's what your favorite theory of content says, that's fine with us. Our proposal can be easily rendered consistent with such a view of content by simply replacing the embedded mental predicates (e.g., "*believe*") with technical terms "bel", "des", "pret", etc. We might then say that the MM produces the belief that *I bel that p* and the belief that *I des that q*; and that at some point further on in development, these beliefs acquire the content *I believe that p, I desire that q*, and so forth. That said, we propose to ignore this subtlety for the rest of the paper.

The core claim of our theory is that the MM is a distinct mechanism that is specialized for detecting one's own mental states.[6] However, it is important to note that on our account of mindreading, the MM is not the *only* mental mechanism that can generate representations with the content *I believe that p*. Representations of this sort can also be generated by ToM. Thus it is possible that in some cases, the ToM and the MM will produce *conflicting* representation of the form *I believe that p*. For instance, if the Theory of Mind is deficient, then in some cases it might produce an inaccurate representation with the content *I believe that p* which conflicts with accurate representations generated by the MM. In these cases, our theory does not specify how the conflict will be resolved or which representation will guide verbal behavior and other actions. On our view, it is an open empirical question how such conflicts will be resolved,

---

[6]As we've presented our theory, the MM is a mechanism that is distinct from the ToM. But it might be claimed that the MM that we postulate is just a *part* of the ToM. Here the crucial question to ask is whether it is a "dissociable" part which could be selectively damaged or selectively spared. If the answer is no, then we think the evidence counts against this view (Nichols & Stich 2002). If the answer is yes (MM is a dissociable part of ToM) then there is nothing of substance left to fight about. That theory is a notational variant of ours.

and this feature of our view will be of some significance for our discussion of the developmental evidence in section 4.

3.2. The Monitoring Mechanism and perceptual states

Of course, the MM Theory is not a complete account of self-awareness. One important limitation is that the MM is proposed as the mechanism underlying self-awareness of one's propositional attitudes, and it's quite likely that the account cannot explain awareness of one's own perceptual states. Perceptual states obviously have phenomenal character, and there is a vigorous debate over whether this phenomenal character is fully captured by a representational account (e.g., Tye 1995, Block forthcoming). If perceptual states can be captured by a representational or propositional account, then perhaps the MM can be extended to explain awareness of one's own perceptual states. For, as noted above, our proposed MM simply copies representations into representation schemas, e.g., it copies representations from the Belief Box into the schema "I believe that ___". However, we're skeptical that perceptual states can be entirely captured by representational accounts, and as a result, we doubt that our MM Theory can adequately explain our awareness of our own perceptual states. Nonetheless, we think it is plausible that some kind of monitoring account (as opposed to a TT account) might apply to awareness of one's own perceptual states. Since it will be important to have a sketch of such a theory on the table, we will provide a brief outline of what the theory might look like.

In specifying the architecture underlying awareness of one's own perceptual states, the first move is to posit a "Percept Box". This device holds the percepts produced by the perceptual processing systems. We propose that the Percept Box feeds into the Belief Box in two ways. First and most obviously, the contents of the Percept Box lead the subject to have beliefs about the world around her, by what we might call a Percept-to-Belief Mediator. For instance, if a normal adult looks into a quarry, her perceptual system will produce percepts that will, *ceteris paribus*, lead her to form the belief that *there are rocks down there*. Something at least roughly similar is presumably true in dogs, birds and frogs. Hence, there is a mechanism (or set of mechanisms) that takes percepts as input and produces beliefs as output. However, there is also, at least in normal adult humans, another way that the Percept Box feeds into the Belief Box – we form beliefs *about our percepts*. For example, when looking into a quarry I might form the belief that *I see rocks*. We also form beliefs about the similarity between percepts – e.g., *this toy rock looks like that real rock*. To explain this range of capacities, we tentatively propose that there is a set of Percept-Monitoring Mechanisms that take input from the Percept Box and produce beliefs about the percepts.[7] We represent this account

[7] How many PMMs are there? A thorough discussion of this is well beyond the scope of this paper, but evidence from neuropsychology indicates that there might be numerous PMMs which can be selectively impaired by different kinds of brain damage. For instance, "achromatopsia" is a condition in which some subjects claim to see only in black and white, but can in fact make some color discriminations. "In cases of

in figure 8. Note that the PMM will presumably be a far more complex mechanism than the MM. For the PMM must take perceptual experiences and produce representations about those perceptual experiences. We have no idea how to characterize this further in terms of cognitive mechanisms, and as a result, we are much less confident about this account than the MM account.

FIGURE 8 ABOUT HERE

4. Developmental evidence and the Theory Theory

The Theory Theory of self-awareness is widely endorsed among researchers working on mindreading, and there are two prominent clusters of arguments offered in support of this account. One of these clusters appeals to evidence on autism as support for a Theory Theory account of self-awareness (Baron-Cohen 1989, Carruthers 1996, Frith & Happé 1999). We consider and reject this cluster of arguments in a companion piece to this paper (Nichols & Stich 2002). However, in this paper, we restrict ourselves to the other cluster of arguments. Perhaps the best known and most widely discussed arguments for the Theory Theory account of self-awareness come from developmental work charting the relation between performance on theory of mind tasks for self and theory of mind tasks for others.[8] In this section we propose to discuss the developmental arguments for and against the TT account of self-awareness.

---

achromatopsia... there is evidence that some aspects of color processing mechanisms continue to function... However... there is no subjective experience of color" (Young 1994, 179). Similarly, prosopagnosiacs claim not to recognize faces; however, many prosopagnosiacs exhibit covert recognition effects in their electrophysiological and behavioral responses (Young 1998, 283-287). Achromatopsia and prosopagnosia are, of course, independent conditions. Prosopagnosiacs typically have no trouble recognizing colors and patients with achromatopsia typically have no trouble recognizing faces. So, it's quite possible that prosopagnosia involves a deficit to a PMM that is not implicated in color recognition and that achromatopsia involves a deficit to a distinct PMM that is not implicated in face recognition. This issue is considerably complicated by the fact that some theorists (e.g., Dennett 1991) maintain that neuropsychological findings like these can be explained by appealing to the mechanisms that build up the multiple layers of the percept itself. We won't treat this complicated issue here. Our point is just that if achromatopsia and prosopagnosia do involve deficits to percept-monitoring mechanisms, it is plausible that they involve deficits to independent PMMs.

[8]The label "theory of mind tasks" is used to characterize a range of experiments that explore the ability to attribute mental states and to predict and explain behavior. For example, as we will discuss later, one prominent theory of mind tasks is the "false belief task". In one version of this task, the subject is shown that a candy box has pencils in it, and the subject has to determine whether a target who has not been shown what is in the box will believe that the box has candy or pencils in it.

It is our contention that the empirical evidence produced by developmental psychologists does not support the TT over our Monitoring Mechanism theory. Rather, we shall argue, in some cases both theories can explain the data about equally well, while in other cases the Monitoring Mechanism theory has a clear advantage over the Theory Theory. Before we present the arguments, it may be useful to provide a brief reminder of the problems we've raised for various versions of the TT account:

> 1. Version 1 looks to be hopelessly implausible; it cannot handle some of the most obvious facts about self-awareness.
>
> 2. Version 2 is a mystery theory; it maintains that there is special source of information exploited in reading one's own mind, but it leaves the source of this additional information unexplained.
>
> 3. Version 3 faces the embarrassment that if information about the representations in the Belief Box & Desire Box is available, then no theory is needed to explain self-awareness; ToM has nothing to do.

We think that these considerations provide an important prima facie case against the Theory Theory account of self-awareness, though we also think that, as in any scientific endeavor, solid empirical evidence might outweigh the prima facie considerations. So we now turn to the empirical arguments.

The Theory Theory predicts that subjects' performance on theory of mind tasks should be about equally good (or equally bad) whether the tasks are about one's own mental states or the mental states of another person. In perhaps the most systematic and interesting argument for the TT, Gopnik & Meltzoff maintain that there are indeed clear and systematic correlations between performance on theory of mind tasks for self and for others (see Table 1, reproduced from Gopnik & Meltzoff 1994, Table 10.1). For instance, Gopnik & Meltzoff note that children succeed at perceptual tasks for themselves and others before the age of 3. Between the ages of 3 and 4, children begin to succeed at desire tasks for self and for other. And at around the age of 4, children begin to succeed at the false belief task for self and for other. "The evidence," Gopnik & Meltzoff maintain,

> 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 false beliefs and source, they do not report that they had those states themselves. Moreover, and in some ways most strikingly, the intermediate case of desire is intermediate for self and other (179-180).

This “extensive parallelism” is taken to show that “our knowledge of ourselves, like our knowledge of others, is the result of a theory” (Gopnik & Meltzoff 1994, 168). Thus the argument purports to establish a broad-based empirical case for the Theory Theory of self-awareness. However, on our view quite the opposite is the case. In the pages to follow we will try to show that the data don’t provide *any* support for the Theory Theory over the Monitoring Mechanism theory that we have proposed, and that some of the data that is comfortably compatible with MM cannot be easily explained by the TT. Defending this claim is rather a long project, but fortunately the data are intrinsically fascinating.

States	Other	Self
Easy		
Pretense	Before age 3 (Flavell et al., 1987)	Before age 3 (Gopnik & Slaughter, 1991)
Imagination	Before age 3 (Wellman & Estes, 1986)	Before age 3 (Gopnik & Slaughter, 1991)
Perception (Level 1)	Before age 3 (Flavell et al. 1981)	Before age 3 (Gopnik & Slaughter, 1991)
Intermediate		
Desire	Age 3-4 (Flavell et al., 1990)	Age 3-4 (Gopnik & Slaughter, 1991)
Difficult		
Sources of belief	After age 4 (O'Neill et al., 1992)	After age 4 (Gopnik & Graf, 1988)
False belief	After age 4 (Wimmer & Perner, 1983)	After age 4 (Gopnik & Astington, 1988)

Table 1: From Gopnik & Meltzoff 1994, 180.

4.1. The parallelism prediction

Before we proceed to the data, it's important to be clear about the structure of Gopnik & Meltzoff's argument and of our counter-argument in favor of the Monitoring Mechanism theory. If Gopnik & Meltzoff are right that there is an "extensive parallelism," that would support the Theory Theory since the Theory Theory *predicts* that there will be parallel performance on parallel theory of mind tasks for self and other. According to the Theory Theory, in order to determine one's own mental states, one must exploit the same Theory of Mind that one uses to determine another's mental states. So, if a child's Theory of Mind is not yet equipped to solve certain third person tasks, then the child should also be unable to solve the parallel first person task.

By contrast, for many of the tasks we'll consider, our theory simply doesn't make a prediction about whether there will be parallel performance on self and other-versions of the tasks. On our theory, the special purpose mechanisms for detecting one's own mental states (MM & PMM), are quite independent from the mechanism for reasoning about mental states and detecting the mental states of others (ToM). Hence, the ability to detect one's own mental states and the ability to detect another's mental states need not

show similar developmental trajectories, though in some cases they might. What our theory does predict is that the capacity to detect one's own mental states, though not necessarily the capacity to reason about them, should emerge quite early, since the theory claims that the MM and the PMM are innate and on line quite early in development. Also, as noted in section 3, our theory allows for the possibility that the ToM *can* be used in attributing mental states to oneself. So it may well turn out that sometimes subjects produce inaccurate self-attributions because they are relying on the ToM. Since our theory provides no *a priori* reason to expect extensive parallel performance in detecting mental states in oneself and others, if there is extensive parallelism our theory would be faced with a major challenge -- it would need to provide some additional and independently plausible explanation for the existence of the parallelism in each case where it is found. But if, as we shall argue, the parallelism is largely illusory, then it is the Theory Theory that faces a major challenge -- it has to provide some plausible explanation for the fact that the parallelism it predicts does not exist.

4.2. Theory Theory meets data

Gopnik & Meltzoff argue for the TT by presenting a wide range of cases in which, they maintain, subjects show parallel performance on self and other versions of theory of mind tasks, and at first glance the range of parallels looks like very impressive indeed. However, we'll argue that on closer inspection, this impression is quite misleading. In some cases, there really is parallel performance, but these cases do not support the TT over our MM theory, since in these cases both theories do about equally well in explaining the facts; in some cases, the evidence for parallel performance is dubious; and in several other cases, there is evidence that performance is *not* parallel. These cases are of particular importance since they are compatible with the MM account and *prima facie* incompatible with the Theory Theory. In the remainder of this section we will consider each of these three classes of cases.

4.2.1. Cases where the parallelism is real

(i) The "easy" tasks

There are a range of tasks that Gopnik & Meltzoff classify as *easy for other and easy for self*. They claim that pretense, imagination, and perception (level 1 perspective taking) are understood for both self and other before age 3. At least on some tasks, this claim of parallel performance seems to be quite right. Simple perceptual tasks provide perhaps the clearest example. Lempers and colleagues (Lempers et al. 1977) found that 2½ year old children succeeded at "level 1" perspective-taking tasks, in which the children had to determine whether another person could see an object or not. For instance, if a young child is shown that a piece of cardboard has a picture of a rabbit on one side and a picture of a turtle on the other, and if the child is then shown the turtle side, the child can correctly answer that the person on the other side of the cardboard sees the picture of the rabbit. Using similar tasks, Gopnik & Slaughter (1991) found that 3-year old children could also successfully report their own past perceptions. As Gopnik &

Meltzoff characterize it, this task is “easy” for other and “easy” for self, and Gopnik & Meltzoff put forward such cases as support for the TT.

As we see it, however, the fact that level-1 perspective-taking tasks are easy for other and for self does not count as evidence for the TT over our MM theory. To see why, let us consider first the *self* case and then the *other* case. On our account, MM is the mechanism responsible for self-awareness of propositional attitudes and, we have tentatively suggested, another mechanism (or family of mechanisms), the Percept-Monitoring Mechanism, underlies awareness of one’s own perceptual states. The PMM, like the MM, is hypothesized to be innate and to come on line quite early in development. Thus the PMM is up and running by the age of 2½, well before ToM is fully in place. So our theory predicts that quite young children should be able to give accurate reports about their own perceptual states. Let's turn now to the *other* case. Both the Theory Theory and our theory maintain that the detection of mental states in others depends on the Theory of Mind and, like Theory Theorists, we think that evidence on visual perspective taking (e.g., Lempers et al. 1977) shows that part of the ToM is on line by the age of 2½. It is of some interest to determine why the part of ToM that subserves these tasks emerges as early as it does, though neither the TT nor our theory currently has any explanation to offer. For both theories it is just a brute empirical fact. So here's the situation: Our theory predicts that awareness of one's own perceptions will emerge early, and has no explanation to offer for why the part of ToM that subserves the detection of perceptual states in others emerges early. By contrast, TT predicts that both self and other abilities will emerge at the same time, but has no explanation to offer for why they both emerge early. By our lights this one is a wash. Neither theory has any clear explanatory advantage over the other.

Much the same reasoning shows that Gopnik & Meltzoff’s cases of pretense and imagination don’t lend any significant support to the Theory Theory over our theory. There is some evidence that by the age of 3, children have some understanding of pretense and imagination in others (e.g., Wellman and Estes 1986), though as we'll see in section 4.2.3, there is also some reason for skepticism. However, whatever the ontogeny is for detecting pretense and imagination in others, the TT account can hardly offer a better explanation than our account, since we simply *are* Theory Theorists about the detection of mental states in others, and neither we nor the Theory Theorists have any explanation to offer for the fact that the relevant part of ToM emerges when it does. As in the case of perception, our theory does have an explanation for the fact that the ability to detect one's own pretenses and imaginings emerges early, since on our view this process is subserved by the MM which is up and running by the age of 2½, but we have no explanation for the fact (if indeed it is a fact) that the part of ToM that subserves the detection of pretenses and imaginings in others also emerges early. The TT, on the other hand, predicts that self and other abilities will both emerge at the same time, but does not explain why they both emerge early. So here, as before, neither theory has any obvious explanatory advantage over the other.

(ii) Sources of belief

A suite of studies by Gopnik, O'Neill and their colleagues (Gopnik & Graf 1988; O'Neill & Gopnik 1991; O'Neill et al. 1992) show that there is a parallel between performance on source tasks for self and source tasks for others. In the self-versions of these tasks, children came to find out which objects were in a drawer either by seeing the object, being told or inferring from a simple cue. After establishing that the child knows what's in the drawer, the child is asked "how do you know that there's an x in the drawer?" This question closely parallels the question used to explore children's understanding of the sources of another's belief (O'Neill et al. 1992). O'Neill and her colleagues found that while 4-year olds tended to succeed at the other-person version of the task, 3- year olds tended to fail it; similarly, Gopnik & Graf (1988) found that 4-year olds tended to succeed at the self-version of the task, but 3 year-olds tended to fail it. For instance, 3-year olds often said that their knowledge came from seeing the object when actually they had been told about the object, and 3-year olds made similar errors when judging the source of another person's knowledge.

These results are interesting and surprising, but they are orthogonal to the issue at hand. The Monitoring Mechanism posited in our theory is a mechanism for *detecting* mental states, not for reasoning about them. But questions about the sources of one's beliefs or knowledge cannot be answered merely by *detecting* one's own mental states. Rather, questions about how you gained knowledge fall into the domain of *reasoning* about mental states, and on our theory that job is performed by the Theory of Mind. So, on our theory, questions about sources will implicate the ToM both for self and other. Hence, our theory, like the Theory Theory, predicts that there will be parallel performance on tasks like the source tasks.

4.2.2. The relevant but dubious data

In Gopnik and Meltzoff's table displaying extensive parallelism, there are two remaining cases that can't be dismissed as irrelevant. However, we will argue that the cases fall far short of clear support for the Theory Theory.

(i) False belief

The false belief task, probably the most famous theory of mind task, was first used by Wimmer and Perner (1983). In their version of the experiment, children watched a puppet show in which the puppet protagonist, Maxi, puts chocolate in a box and then goes out to play. While Maxi is out, his puppet mother moves the chocolate to the cupboard. When Maxi returns, the children are asked where Maxi will look for the chocolate. Numerous studies have now found that 3-year old children typically fail tasks like this, while 4 year olds typically succeed at them (e.g., Baron-Cohen et al. 1985, Perner et al. 1987). This robust result has been widely interpreted to show that the ToM (or some quite fundamental component of it) is not yet in place until about the age of 4.

On closely matched tasks, Gopnik & Astington (1988) found a correlation between failing the false belief task for another and failing it for oneself. Gopnik &

Astington (1988) presented children with a candy box and then let the children see that there were really pencils in the box. Children were asked "what will Nicky think is in the box?" and then "When you first saw the box, before we opened it, what did you think was inside it?". Children's ability to answer the question for self was significantly correlated with their ability to answer the question for other. Thus, here we have a surprising instance of parallel performance on tasks for self and other.[9] This is, of course, just the outcome that the Theory Theory would predict. For the Theory Theory maintains that ToM is crucial both in the detection of other people's beliefs and in the detection of one's own. Thus if a child's ToM has not yet developed to the point where it can detect other people's beliefs in a given situation, it is to be expected that the child will also be unable to detect her own beliefs in that context. And this, it appears, is just what the experimental results show.

What about our theory? What explanation can it offer for these results? The first step in answering this question is to note that in the *self* version of the false belief task, the child is not actually being asked to report on her *current* belief, but rather to recall a belief she had in the recent past. Where might such memories come from? The most natural answer, for a theory like ours, is that when the child first sees the box she believes that there is candy in it, and the MM produces a belief with the content *I believe that there is candy in the box*. As the experiment continues and time passes that belief is converted into a past tense belief whose content is (roughly) *I believed that there was candy in the box*. But, of course, if that were the end of the story, it would be bad news for our theory, since when asked what she believed when she first saw the box the child reports that she believes *that there were pencils in the box*. Fortunately, that is *not* the end of the story. For, as we noted in section 3, in our theory MM is not the only mechanism capable of generating beliefs with the content *I believe(d) that p*. ToM is also capable of producing such beliefs, and sometimes ToM may produce a belief of that form that will conflict with a belief produced by MM. That, we propose, is exactly what is happening in the Gopnik and Astington experiment when younger children fail to report their own earlier false belief. As the results in the other-version of the task indicate, the ToM in younger children has a strong tendency to attribute beliefs that the child actually believes to be true. So when asked what she believed at the beginning of the experiment, ToM mistaken concludes that *I believed that there were pencils in the box*.[10] Thus, on our account, there will be two competing and incompatible representations in the child's Belief Box. And to explain the fact that the child usually relies on the mistaken ToM generated belief, rather than on the correct MM generated belief, we must suppose that the memory trace is relatively weak, and that when the child's cognitive system has to

---

[9]Similarly, Baron-Cohen 1991 found that in autism, there are correlations between failing the false belief task for other and failing the task for self.

[10] Some theorists, most prominently Fodor (1992), have explained the results in the other-version of the task by claiming that young children do not use the ToM in these tasks. They arrive at their answer, Fodor argues, by using a separate reality biased strategy. We need take no stand on this issue, since if Fodor is correct then it is plausible to suppose that the same reality biased strategy generates a mistaken *I believed that there were pencils in the box* representation in the self-version of the task.

decide which belief about her past belief to rely on, the MM generated memory trace typically loses.

At this point, we suspect, a critic might protest that this is a singularly unconvincing explanation. There is, the critic will insist, no reason to think that the MM generated memory will typically be weaker than the ToM generated belief; it is just an *ad hoc* assumption that is required to get our theory to square with the facts. And if this were the end of the story, the critic would be right. Fortunately for us, however, this is not the end of the story. For there is evidence that provides independent support for our explanation and undercuts the TT account. Recent work by German and Leslie exploring performance on self- and other- versions of the false belief task indicates that *if memory enhancements are provided, young children's performance on self-versions improves, while their performance on other-versions stays about the same.* German & Leslie devised a task in which a child would hide a biscuit and then search for it in the wrong place, because it had been moved when the child was behind a screen. In one condition, the child was then shown a videotape of the entire sequence of events -- hiding, moving and searching -- and asked, at the appropriate point, "Why are you looking there?" and then "When you were looking for the biscuit, where did you think the biscuit was?" In another condition, after the same hiding, moving and searching sequence, the videotape was "accidentally" rewound too far, and the child watched another child in an identical situation. At the appropriate point, the child was asked "Why was she looking there?" and "When she was looking for the biscuit, where did she think the biscuit was?" German & Leslie found that children who were shown their own mistaken search were much more likely to offer a false belief explanation and to attribute a false belief than were children who were shown another's mistaken search (German & Leslie, forthcoming). This fits nicely with our proposed explanation for why young children fail the false belief task for the self. However, it's difficult to see how a Theory Theorist could explain these results. For according to the Theory Theory, if the child has a defective ToM, the child should make the same mistakes for himself that he does for another. If there is no MM to generate a correct belief which becomes a correct memory, then giving memory enhancements should not produce differential improvement.

(ii) Desire

Another source of data that might offer support to the TT comes from work on understanding desires. Gopnik & Meltzoff maintain that 3-year olds are just beginning to understand desire in others, and Gopnik & Slaughter found that a significant percentage of children make mistakes about their own immediately past desires. The Gopnik & Slaughter own-desire tasks were quite ingenious. In one of the tasks, they went to a daycare center just before snack time and asked the child whether he was hungry. The hungry child said "yes" and proceeded to eat all the snack he desired. Then the experimenter asked, "When I first asked you, before we had the snack, were you hungry then?" (102). Gopnik & Slaughter found that 30-40% of the 3 year olds mistakenly claimed that they were in their current desire state all along. This surprising result is claimed to parallel Flavell et al.'s (1990) finding that a significant percentage of 3 year

olds make mistakes on desire tasks for others. In the Flavell tasks, the child observes Ellie make a disgusted look after tasting a cookie, and the child is asked "Does Ellie think it is a yummy tasting cookie?" (Flavell et al. 1990, 918). Gopnik & Meltzoff remark that the "absolute levels of performance were strikingly similar" to the results reported by Flavell et al. (Gopnik & Meltzoff 1994, 179), and they cite this as support for the parallel performance hypothesis.

The central problem with this putative parallel is that it's not at all clear that the tasks are truly parallel tasks. In Gopnik & Slaughter's tasks, 3 year olds are asked about a desire that they don't currently have because it was recently satisfied. It would be of considerable interest to couple Gopnik & Slaughter's own-desire version of the hunger task with a closely matched other-person version of the task. For instance, the experiment could have a satiated child watch another child beginning to eat at snack time and ask the satiated child, "Is he hungry?" If the findings on this task paralleled findings on the own-desire version, that would indeed be an important parallel. Unfortunately, the putatively parallel task in Flavell et al. that Gopnik & Meltzoff cite is quite different from the Gopnik & Slaughter task. In the Flavell tasks, the child is asked whether the target thinks the cookie is "yummy tasting" (Flavell et al. 1990, 918). The task doesn't explicitly ask about desires at all. Flavell and his colleagues themselves characterize the task as exploring children's ability to attribute value *beliefs*. Further, unlike the Gopnik & Slaughter task, the Flavell et al. tasks depend on expressions of disgust. Indeed, there are so many differences between these tasks that we think it's impossible to draw any conclusions from the comparison.

In this section we have considered the best cases for the Theory Theory, and it is our contention that the data we've discussed don't provide much of an argument for the Theory Theory. For there are serious empirical problems with both cases, and even if we ignore these problems, the data certainly don't establish the "extensive parallelism" that is predicted by the Theory Theory. Moreover, there are results not discussed by Gopnik and Meltzoff which, we think, strongly suggest that the parallelism on which their argument depends simply does not exist.

4.2.3. Evidence against the self-other parallelism

In this section we will review a range of data indicating that often there is *not* a parallel between performance on self and other versions of theory of mind tasks. We are inclined to think that these data completely uproot Gopnik & Meltzoff's parallelism argument, and constitute a major challenge to the Theory Theory of self-awareness itself.

(i) Knowledge *vs*. ignorance

In knowledge versus ignorance experiments, Wimmer and colleagues found a significant difference between performance on closely matched tasks for self and other. (Wimmer et al. 1988). After letting children in 2 conditions either look in a box or not

look in a box, the researchers asked them "Do you know what is in the box or do you not know that?" The 3 year olds performed quite well on this task. For the other-person version of the task, they observed another who either looked or didn't look into a box. They were then asked: "Does [name of child] know what is in the box or does she [he] not know that?" (1988, 383). Despite the almost verbatim similarity between this question and the self-version, the children did significantly worse on the other-version of this question (see also Nichols 1993). Hence, we have one case in which there is a significant *difference* between performance on a theory of mind task for self and performance on the task for other. And there's more to come.

(ii) Pretense and imagination

Gopnik & Meltzoff maintain that children under age 3 understand pretense for others and for self. Although there are tasks on which young children exhibit some understanding of pretense (e.g., Wellman & Estes 1986), the issue has turned out to be considerably more complicated. It's clear from the literature on pretend play that from a young age, children are capable of reporting their own pretenses. Indeed, Gopnik & Slaughter (1991) show that 3 year old children can easily answer questions about their past pretenses and imaginings. Despite this facility with their own pretenses, it doesn't seem that young children have an adequate theory of pretense (Lillard 1993, Nichols & Stich 2000). For instance, Lillard's (1993) results suggests that children as old as four years think that someone can pretend to be a rabbit without knowing anything about rabbits. More importantly for present purposes, although young children have no trouble detecting and reporting their own pretenses (e.g., Leslie 1994a), children seem to be significantly worse at recognizing pretense in others (Flavell et al 1987; Rosen et al. 1997). Indeed, recent results from Rosen et al. (1997) indicate that young children have a great deal of difficulty characterizing the pretenses of others. Rosen and his colleagues had subjects watch a television show in which the characters were sitting on a bench but pretending to be on an airplane. The researchers asked the children: "Now we're going to talk about what everyone on *Barney* is thinking about. Are they thinking about being on an airplane or about sitting on a bench outside their school?" (1135). They found that 90% of the 3 year olds answered incorrectly that everyone was thinking about sitting on a bench. By contrast, in Gopnik & Slaughter's experiments, 3 year old children did quite well on questions about what they themselves were pretending or imagining. In one of their pretense tasks, the child was asked to pretend that an empty glass had orange juice in it; the glass was turned over, and the child was subsequently asked to pretend that it had hot chocolate in it. The child was then asked, "When I first asked you…. What did you pretend was in the glass then?" (Gopnik & Slaughter 1991, 106). Children performed near ceiling on this task. In Gopnik & Slaughter's imagination task, the children were told to close their eyes and think of a blue doggie, then they were told to close their eyes and think of a red balloon. The children were then asked, "When I first asked you…., what did you think of then? Did you think of a blue doggie or did you think of a red balloon?" (G&S 1991, 106). Over 80% of the 3 year olds answered this correctly. Although the Gopnik & Slaughter pretense and imagination tasks aren't exact matches for the Rosen et al. task, the huge difference in the results suggests that children do much

better on pretense and imagination tasks for self than they do on pretense and imagination tasks for another person. Hence, it seems likely that children can detect and report their own pretenses and imaginings before they have the theoretical resources to detect and characterize pretenses and imaginings in others.

(iii) Perspective taking

As we noted earlier, children as young as 2½ years are able to succeed at "level 1" perspective taking tasks both for others and for themselves. However, there is a cluster of more difficult perspective taking tasks, "level 2" tasks, in which young children do significantly better in the self-version than in the other-version . These tasks require the child to figure our how an object looks from a perspective that is different from her own current perspective. In one task, for example, the child is shown a drawing of a turtle that looks to be lying on his back when viewed from one position and standing on his feet when viewed from another position. The child is asked whether the turtle is on its back or on its feet; then the child is asked how the person across the table sees the turtle, on its back or on its feet. Children typically don't succeed at these tasks until about the age of 4. However, contrary to the parallel performance hypothesis, Gopnik & Slaughter (1991) found that 3 year olds did well on a self-version of the task. They had the child look at the drawing of the turtle and then had the child change seats with the experimenter. The child was subsequently asked "When I first asked you, before we traded seats, how did you see the turtle then, lying on his back or standing on his feet" (106). Gopnik & Slaughter were surprised at how well the 3 year olds did on this task. They write, "Perhaps the most surprising finding was that performance on the level 2 perception task turned out to be quite good, and was not significantly different from performance on the pretend task. Seventy-five percent of the 3-year-olds succeeded at this task, a much higher level of performance than the 33% to 50% reported by Masangkay et al. (1974) in the other person version of this task" (Gopnik & Slaughter 1991, 107). Here, then, is another example of a theory of mind task in which the self-version of the task is significantly easier for subjects than the other-version of the task. So we have yet another case in which the Theory Theory's prediction of extensive parallelism is disconfirmed. [11]

[11]Gopnik & Meltzoff have also produced results that suggest a disparity between performance on self- and other-versions of a very simple perspective taking task. They found that when 24 month olds were asked to hide an object from the experimenter, they "consistently hid the object egocentrically, either placing it on the experimenter's side of the screen or holding it to themselves so that neither they nor the experimenter could see it" (reported in Gopnik & Meltzoff 1997, 116). Given that Gopnik & Meltzoff characterize the child's performance as "egocentric", it seems quite likely that the children would succeed at versions of this task that asked the child to hide the object from herself. Hence, one expects that children would perform significantly better on a self-version of the task than on the other-version of the task. If in fact the 2 year old child can't solve the hiding task for another person, but can solve it for self, then this looks like another case that counts against the extensive parallelism predicted by the Theory Theory.

4.3. What conclusions can we draw from the developmental data?

We now want to step back from the details of the data to assess their implications for the debate between the Theory Theory and our Monitoring Mechanism Theory. To begin, let's recall what each theory predicts, and why. The TT maintains that the ToM is centrally involved in detecting and reasoning about both one's own mental states and other people's. But the TT makes no claims about when in the course of development various components of ToM are acquired or come on line. Thus TT makes no predictions about when specific mind reading skills will emerge, but it does predict that any given mind reading skill will appear at about the same time in self and other cases. MM, by contrast, maintains that ToM is involved in detecting and reasoning about other people's mental states and in reasoning about one's own mental states, but that a separate Monitoring Mechanism (or a cluster of such mechanisms) is typically involved when we detect our own mental states. MM also claims that the Monitoring Mechanism(s) come on line quite early in development. Thus MM predicts that children will be able to detect (but not necessarily reason about) their own mental states quite early in development. But it does not predict any particular pattern of correlation between the emergence of the capacity to detect one's own mental states and the emergence of the capacity to detect other people's mental states.

Which theory does better at handling the data we have reviewed? As we see it, the answer if clear: MM is compatible with all the data we have reviewed, while some of the data is seriously problematic for the TT. To make the point as clearly as possible, let's assemble a list of the various mind reading phenomena we have reviewed:

1) Level 1 perspective taking. This emerges early for both self and other. TT predicts the parallel emergence and is compatible with, but does not predict, the early emergence. MM predicts the early emergence in the self case and is compatible with but does not predict the early emergence in the other case. Neither theory has an advantage over the other.

2) Pretense and imagination. It is clear that self detection emerges early, as MM predicts. However, there is some recent evidence indicating that detection and understanding of pretense in others does not emerge until much later. If this is right, it is a problem for TT, though not for MM.

3) Sources of belief: The ability to identify sources of belief emerges at about the age of 4 in both the self and the other case. Since this is a reasoning problem not a detection problem, both theories make the same prediction.

4) False belief: Recent evidence indicates that if memory enhancements are provided, young children do better on the self-version of false belief tasks than on the other-version. This is compatible with MM but quite problematic for TT.

5) Desire: The evidence available does not use well matched tasks, so no conclusions can be drawn about either TT or MM.

6) Knowledge *vs*. ignorance: 3 year olds do much better on the self-version than on the other-version. This is compatible with MM but problematic for TT.

7) Level 2 perspective taking: Here again, 3 year olds do better on the self-version than on the other-version, which is a problem of TT but not for MM.

Obviously, the extensive parallelism between self and other cases on which Gopnik and Meltzoff rest their case for the Theory Theory of self-awareness is not supported by the data. Conceivably a resourceful Theory Theorist could offer plausible explanations for each of the cases in which the parallel predicted by TT breaks down. But in the absence of a systematic attempt to provide such explanations we think it is clear that the developmental evidence favors our theory of self-awareness over the Theory Theory.

5. Conclusion

The empirical work on mindreading provides an invaluable resource for characterizing the cognitive mechanisms underlying our capacity for self-awareness. However, we think that developmental psychologists have drawn the wrong conclusions from the data. Contrary to the claims of Theory Theorists, the developmental evidence indicates that the capacity for self-awareness does not depend on the Theory of Mind. It's much more plausible, we have argued, to suppose that self-awareness derives from a Monitoring Mechanism that is independent of the Theory of Mind. Hence, we think that at this juncture in cognitive science, the most plausible account of self-awareness is that the mind comes pre-packaged with a set of special-purpose mechanisms for reading one's own mind.

Acknowledgements: We would like to thank Peter Carruthers, Catherine Driscoll, Luc Faucher, Trisha Folds-Bennett, Gary Gates, Rochel Gelman, Alison Gopnik, Alan Leslie, Brian Loar, Dominic Murphy, Brian Scholl, Eric Schwitzgebel, and Robert Woolfolk for discussion and comments on earlier drafts of this paper. Earlier versions of this paper were presented at a conference sponsored by the Center for Philosophical Education in Santa Barbara, California, at the Rutgers University Center for Cognitive Science, and at the Institute for the Study of Child Development, Robert Wood Johnson Medical School. We are grateful for the constructive feedback offered by members of the audience on all of these occasions.

References:

Armstrong, D. (1968). *A materialist theory of the mind*. London: Routledge & Kegan Paul.

Baron-Cohen, S. (1989). Are autistic children 'behaviorists'?. *Journal of Autism and Developmental Disorders*, 19, 579-600.

Baron-Cohen, S. (1991). The development of a theory of mind in autism: deviance and delay?. *Psychiatric Clinics of North America*, 14, 33-51.

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

Block, N. (forthcoming). Mental paint.

Carruthers, P. (1996). Autism as mind-blindness: An elaboration and partial defence. In: *Theories of theories of mind,* ed. P. Carruthers & P. Smith. Cambridge: Cambridge University Press.

Dennett, D. (1991). *Consciousness explained*. Boston, MA: Little Brown.

Ericsson, K. & Simon, H. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Flavell, J., Everett, B., Croft, K. & Flavell, E. (1981). Young children's knowledge about visual perception. *Developmental Psychology*, 17, 99-103.

Flavell, J., Flavell, E. Green, F. (1986). Young children's knowledge about the apparent-real and pretend-real distinctions. *Developmental Psychology,* 23, 816-22.

Flavell, J., Flavell, E., Green, F., and Moses, L. (1990). Young children's understanding of fact beliefs versus value beliefs. *Child Development*, 61, 915-928.

Fodor, J. (1992). A theory of the child's theory of mind. *Cognition*, 44, 283-96.

Frith, C. (1994). Theory of mind in schizophrenia. In: *The Neuropsychology of Schizophrenia*, ed. A. David & J. Cutting. Hillsdale, NJ: LEA.

Frith, U. & Happé, F. (1999). Theory of mind and self consciousness: What is it like to be autistic? *Mind & Language*, 14, 1-22.

German, T. & Leslie, A., (forthcoming). Self-other differences in false belief: Recall versus reconstruction.

Goldman, A. (1993a). *Philosophical applications of cognitive science*. Boulder, CO: Westview Press.

Goldman, A. (1993b). The psychology of folk psychology. *Behavioral and Brain Sciences*, 16, 15-28, 101-113.

Goldman, A. (1997). Science, publicity, and consciousness. *Philosophy of Science*, 64, 525-546.

Goldman, A. (2000). The mentalizing folk. In D. Sperber (ed.) *Metarepresentation*. Oxford: Oxford University Press.

Goodman, N. (1983). *Fact, fiction & forecast*, 4th edition. Cambridge, MA: Harvard University Press.

Gopnik, A. (1993). How we know our own minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences*, 16, 1-14.

Gopnik, A. & Astington, J. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child Development*, 59, 26-37.

Gopnik, A. & Graf, P. (1988). Knowing how you know: Young children's ability to identify and remember the sources of their beliefs. *Child Development*, 59, 1366-71.

Gopnik, A. & Meltzoff, A. (1994). Minds, bodies, and persons: Young children's understanding of the self and others as reflected in imitation and theory of mind research. In *Self-awareness in animals and humans*, ed. S. Parker, R. Mitchell, and M. Boccia. New York: Cambridge University Press.

Gopnik, A. & Meltzoff, A. (1997). *Words, thoughts and theories*. Cambridge, MA: MIT Press.

Gopnik, A. & Slaughter, V. (1991). Young children's understanding of changes in their mental states. *Child Development*, 62, 98-110.

Gopnik, A. & Wellman, H. (1994). The theory theory. In S. Gelman & L. Hirschfeld (eds.) *Mapping the Mind*. Cambridge: Cambridge University Press.

Gordon, R. (1995). Simulation without introspection or inference from me to you. In: *Mental Simulation: Evaluations and Applications*, ed. T. Stone and M. Davies. Oxford: Blackwell.

Gordon, R. (1996). Radical simulationism. In: *Theories of Theories of Mind*, ed. P. Carruthers & P. Smith. Cambridge: Cambridge University Press, 11-21.

Lempers, J., Flavell, E., and Flavell, J. (1977). The development in very young children of tacit knowledge concerning visual perception. *Genetic Psychology Monographs*, 95, 3-53.

Leslie, A. (1994a). Pretending and believing: Issues in the theory of ToMM. *Cognition*, 50, 211-238.

Leslie, A. (1994b). ToMM, ToBY and Agency: Core architecture and domain specificity. In L. Hirschfeld & S. Gelman (eds.) *Mapping the mind*. Cambridge: Cambridge University Press, 119-148.

Lillard, A. (1993). Young children's conceptualization of pretense: Action or mental representational state? *Child Development*, 64, 372-386.

Lycan, W. (1987). *Consciousness*. Cambridge, MA: MIT Press.

Masangkay, Z., McCluskey, K., McIntyre, C., Sims-Knight, J., Vaughan, B., & Flavell, J. (1974). The early development of inferences about the visual percepts of others. *Child Development*, 45, 357-366.

McLaughlin, B. & Tye, M. (1998). Is content externalism compatible with privileged access? *Philosophical Review*, 107, 349-80.

Nichols, S. (1993). Developmental evidence and introspection. *Behavioral and Brain Sciences*, v. 16, no. 1, 64-65.

Nichols, S. and Stich, S. (1998). Rethinking co-cognition. *Mind & Language*, 13, 499-512.

Nichols, S. and Stich, S. (2000). A cognitive theory of pretense. *Cognition*, 74, 115-147.

Nichols, S. and Stich, S. (2002). How to Read Your Own Mind: A Cognitive Theory of Self-Consciousness. *Consciousness: New Philosophical Essays*, eds. Q. Smith and P. Jokic. Oxford: Oxford University Press.

Nichols, S. and Stich, S. (forthcoming). *Mindreading*. Oxford: Oxford University Press.

Nichols, S., Stich, S., and Leslie, A. (1995). Choice effects and the ineffectiveness of simulation: Response to Kuhberger et al.. *Mind & Language*, 10, 437-445.

Nichols, S., Stich, S., Leslie, A., and Klein, D. (1996). Varieties of off-line simulation. *Theories of Theories of Mind*, eds. P. Carruthers and P. Smith. Cambridge: Cambridge University Press, 39-74.

Nisbett, R. and Schacter, S. (1966). Cognitive manipulation of pain. *Journal of Experimental Social Psychology*, 21, 227-236.

Nisbett, R. and Wilson, T. (1977). Telling more than we can know. *Psychological Review*, 84, 231-59.

O'Neill, D. & Gopnik, A. (1991). Young children's understanding of the sources of their beliefs. *Developmental Psychology*, 27, 390-397.

O'Neill, D., Astington, J. and Flavell, J. (1992). Young children's understanding of the role that sensory experiences play in knowledge acquisition. *Child Development*, 63, 474-91

Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: MIT Press.

Perner, J., Leekam, S. and Wimmer, H. (1987). Three-year olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Experimental Child Psychology*, 39, 437-71.

Putnam, H. (1975). The meaning of meaning. In *Mind, language and reality: Philosophical papers*, vol. 2. Cambridge: Cambridge University Press.

Rosen, C., Schwebel, D, & Singer, J. (1997). Preschoolers' attributions of mental states in pretense. *Child Development*, 68, 1133-1142.

Sellars, W. (1956). Empiricism and the philosophy of mind. *Minnesota studies in the philosophy of science*, vol. 1. University of Minnesota Press. Reprinted in Sellars (1963) *Science, perception and reality*. London: Routledge & Kegan Paul.

Stich, S. (1992). What is a theory of mental representation? *Mind*, 101, 243-261.

Stich, S. (1996). *Deconstructing the Mind.* New York: Oxford University Press.

Stich, S. and Nichols, S. (1992). Folk psychology: Simulation or tacit theory". *Mind & Language*, v. 7, no. 1, 35-71.

Stich, S. and Nichols, S. (1995). Second thoughts on simulation. In: *Mental Simulation: Evaluations and Applications*, ed. A. Stone and M. Davies. Oxford: Basil Blackwell, 87-108.

Stich, S. and Ravenscroft, I. (1994). What *is* folk psychology?" *Cognition*, 50, 1-3, 447-468.

Stich, S. and Nichols, S. (1997). "Cognitive penetrability, rationality, and restricted simulation". *Mind & Language*, 12, 297-326.

Stich, S. and Nichols, S. (1998). Theory theory to the max: A critical notice of Gopnik & Meltzoff's *Words, thoughts, and theories. Mind & Language*, 13, 421-449.

Tye, M. (1995). *Ten problems of consciousness: A representational theory of the phenomenal mind*. Cambridge, MA: MIT Press.

Wellman, H. & Estes, D. (1986). Early understanding of mental entities: A reexamination of childhood realism. *Child Development*, 57, 910-23.

Wimmer, H. & Hartl, M. (1991). The Cartesian view and the theory view of mind: Developmental evidence from understanding false belief in self and other. *British Journal of Developmental Psychology*, 9, 125-28.

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

Wimmer, H., Hogrefe, G., & Perner, J. (1988). Children's understanding of informational access as a source of knowledge. *Child Development*, 59, 386-96.

Young, A. (1994). Neuropsychology of Awareness. In *Consciousness in Philosophy and Cognitive Neuroscience*, ed. A. Revonsuo & M. Kamppinen. Hillsdale, NJ: LEA.

Young, A. (1998). *Face and mind*. Oxford: Oxford University Press.