This article appeared in *Consciousness: New Essays*, eds. Q. Smith and A. Jokic. Oxford University Press, 2003.

How to Read Your Own Mind: A Cognitive Theory of Self-Consciousness

Shaun Nichols & Stephen Stich

1. Introduction

The topic of self-awareness has an impressive philosophical pedigree, and sustained discussion of the topic goes back at least to Descartes. More recently, self-awareness has become a lively issue in the cognitive sciences, thanks largely to the emerging 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, 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 includes 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 arguments from psychopathologies and we will argue that none of the evidence favors the TT over our Monitoring Mechanism Theory.[1] Indeed, in the fifth

[1] Elsewhere, we consider the evidence from development (Nichols & Stich forthcoming a, b). Nichols & Stich (forthcoming b) is intended as a companion piece to this article. In that article, we argue that a closer inspection of the developmental evidence shows that the developmental argument for Theory Theory is unworkable and that the evidence actually poses a problem for the Theory Theory. Of necessity, there is considerable overlap between the present paper and Nichols & Stich (forthcoming b). In both papers, we consider whether the evidence favors the Theory Theory or the Monitoring

section, we will exploit evidence on psychopathologies to provide an argument in favor of the Monitoring Mechanism Theory. On our account, but not on the TT, it is possible for the mechanisms subserving self-awareness and reading other people's minds to be damaged independently. And, we will suggest, this may well be just what is happening in certain cases of schizophrenia and autism. After making our case against the TT and in favor of our theory, we will consider two other theories of self-awareness to be found in the recent literature. The first of these, discussed in section 6, is Robert Gordon's "ascent routine" account (Gordon 1995, 1996), which, we will argue, is clearly inadequate to explain the full range of self-awareness phenomena. The second is Alvin Goldman's (1993a, 1993b, 1997, forthcoming) phenomenological account which, we maintain, is also under-described and admits of two importantly different interpretations. On both of the interpretations, we'll argue, the theory is singularly implausible. But before we do any of this, there is a good deal of background that needs to be set in place.

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)

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).

Mechanism theory, and the theoretical background against which the arguments are developed is largely the same in both papers. As a result, readers familiar with Nichols & Stich (forthcoming b) might skip ahead to section 4.

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)[2]. 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

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

[2]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.

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.[3]

---

[3]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 distinctions (Stich & Nichols 1997; Nichols & Stich 1998). In the first five sections of 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

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 1994) 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 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:

---

accommodate the facts about self-awareness. So, until section 6, we will simply assume that reasoning about other minds depends on a theory.

> 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. Using "think aloud" procedures, researchers have been able to corroborate self-reports of current mental states against other measures. In typical experiments, subjects are given logical or mathematical problems to solve and are instructed to "think aloud" while they work the problems.[4] For instance, people are

[4]To give an idea of how this works, here is an excerpt from Ericsson & Simon's instructions to subjects in think-aloud experiments:

> In this experiment we are interested in what you think about when you find answers to some questions that I am going to ask you to answer. In order to do this I am going to ask you to THINK ALOUD as you work on the problem given. What I mean by think aloud is that I want you to tell me EVERYTHING you are thinking from the time you first see the question until you give an answer (Ericsson & Simon 1993, 378).

asked to think aloud while multiplying 36 times 24 (Ericsson & Simon 1993, 346-7). Subjects' responses can then be correlated with formal analyses of how to solve the problem, and the subject's answer can be compared against the real answer. If the subject's think-aloud protocol conforms to the formal task analysis, that provides good reason to think that the subject's report of his thoughts is accurate (Ericsson & Simon 1993, 330). In addition to these concurrent reports, researchers have also explored retrospective reports of one's own problem solving[5]. For instance Ericsson & Simon discuss a study by Hamilton & Sanford in which subjects were presented with two different letters (e.g., R-P) and asked whether the letters were in alphabetical order. Subjects were then asked to say how they solved the problem. Subjects reported bringing to mind strings of letters in alphabetical order (e.g., LMNOPQRST), and reaction times taken during the problem solving correlated with the number of letters subjects recollected (Ericsson & Simon 1993, 191-192).

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

---

[5]For retrospective reports, immediately after the subject completes the problem, the subject is given instructions like the following:

> Now I want to see how much you can remember about what you were thinking from the time you read the question until you gave the answer. We are interested in what you actually can REMEMBER rather than what you think you must have thought. If possible I would like you to tell about your memories in the sequence in which they occurred while working on the question. Please tell me if you are uncertain about any of your memories. I don't want you to work on solving the problem again, just report all that you can remember thinking about when answering the question. Now tell me what you remember (Ericsson & Simon 1993, 378).

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.[6] 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

---

[6]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.

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 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.[7] 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.

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

---

[7]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'll argue against this view in section 5. 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.

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.[8] We represent this account 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

[8] 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 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.

4. Autism and the Theory Theory

The Theory Theory of self-awareness is widely held among researchers working on mindreading, and there are two prominent clusters of arguments offered in support of this account. One widely discussed set of arguments comes from developmental work charting the relation between performance on theory of mind tasks for self and theory of mind tasks for others.[9] This is an important set of arguments, but an adequate treatment of these arguments requires a close and lengthy inspection of the developmental data, a task we take up in the companion piece to this article (Nichols & Stich forthcoming b). Here we focus on the other central cluster of arguments for the Theory Theory of self awareness. Several prominent advocates of TT have appealed to evidence on autism as support for a Theory Theory account of self-awareness (Baron-Cohen 1989, Carruthers 1996, Frith & Happé 1999). On our view, however, the evidence from autism provides no support at all for this theory of self-awareness. But before we plunge into this debate, 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.

To explicate the arguments based on evidence from autism, we first need to provide a bit of background to explain why data about autism are relevant to the issue of self-awareness. Studies of people with autism have loomed large in the literature in mindreading ever since Baron-Cohen, Leslie & Frith (1985) reported some now famous results on the performance of autistic individuals on the false belief task. The original version of the false belief task was developed by Wimmer & 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,

[9]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".

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. Baron-Cohen and colleagues compared performance on false belief tasks in normal children, autistic children and children with Down syndrome. What they found was that autistic subjects with a mean chronological age of about 12 and mean verbal and non verbal mental ages of 9;3 and 5;5 respectively failed the false belief task (Baron-Cohen et al. 1985). These subjects answered the way normal 3 year olds do. By contrast, the control group of Downs syndrome subjects matched for mental age performed quite well on the false belief task. One widely accepted interpretation of these results is that *autistic individuals lack a properly functioning ToM*.

If we assume that this is correct, then, since the TT account of self-awareness claims that ToM is implicated in the formation of beliefs about one's own mental states, the TT predicts that autistic individuals should have deficits in this domain as well. If people with autism lack a properly functioning ToM and a ToM is required for self-awareness, then autistic individuals should be unable to form beliefs about their own beliefs and other mental states. In recent papers both Carruthers (1996) and Frith & Happé (1999) have maintained that autistic individuals do indeed lack self-awareness, and that this supports the TT account. In this section we will consider three different arguments from the data on autism. One argument depends on evidence that autistic children have difficulty with the appearance/reality distinction. A second argument appeals to introspective reports of adults with Asperger's Syndrome (autistic individuals with near normal IQs), and a third, related, argument draws on autobiographical testimony of people with autism and Asperger's Syndrome.

4.1. Autism and the appearance/reality distinction

Both Baron-Cohen (1989) and Carruthers (1996) maintain that the performance of autistic children on appearance/reality tasks provides support for the view that autistic children lack self-awareness, and hence provides evidence for the TT. The relevant studies were carried out by Baron-Cohen (1989), based on the appearance/reality tasks devised by Flavell and his colleagues. Using those tasks, Flavell and his colleagues found that children have difficulty with the appearance/reality distinction until about the age of four (Flavell et al. 1986). For instance, after playing with a sponge that visually resembles a piece of granite (a "Hollywood rock"), most three year olds claim that the object both is a sponge and looks like a sponge. Baron-Cohen found that autistic subjects also have difficulty with the appearance/reality distinction. When they were allowed to examine a piece of fake chocolate made out of plastic, for example, they thought that the object both looked like chocolate and really was chocolate. "In those tasks that included plastic food," Baron-Cohen reports, "the autistic children alone persisted in trying to eat the object long after discovering its plastic quality. Indeed, so clear was this perseverative behavior that the experimenter could only terminate it by taking the plastic object out of their mouths" (Baron-Cohen 1989, 594).

Though we find Baron-Cohen and Flavell et al.'s work on the appearance/reality distinction intriguing, we are deeply puzzled by the suggestion that the studies done with autistic subjects provide support for the TT account of self-awareness. And, unfortunately, those who think that these studies do support the TT have never offered a detailed statement of how the argument is supposed to go. At best they have provided brief hints like the following:

> [T]he mind-blindness theory would predict that autistic people will lack adequate access to their own experiences as such …, and hence that they should have difficulty in negotiating the contrast between *experience* (appearance) and *what it is an experience of* (reality) (Carruthers 1996, 260-1).

> [Three year old children] appear unable to represent both an object's real and apparent identities simultaneously…. Gopnik and Astington (1988) argued that this is also an indication of the 3-year old's inability to represent the distinction between their representation of the object (its appearance) and their knowledge about it (its real identity). In this sense, the A-R distinction is a test of the ability to attribute mental states to oneself (Baron-Cohen 1989, 591).

> The prediction that this would be an area of difficulty for autistic subjects was supported, and this suggests that these children … are unaware of the A-R distinction, and by implication unaware of their own mental states. These results suggest that when perceptual information contradicts one's own knowledge about the world, the autistic child is unable to separate these, and the perceptual information overrides other representations of an object" (Baron-Cohen 1989, 595)

How might these hints be unpacked? What we have labeled *Argument I* is our best shot at making explicit what Carruthers and Baron-Cohen might have had in mind. Though we are not confident that this is the right interpretation of their suggestion, it is the most charitable reading we've been able to construct. If this *isn't* what they had in mind (or close to it) then we really haven't a clue about how the argument is supposed to work.

Argument I

If the TT of self-awareness is correct then ToM plays a crucial role in forming beliefs about one's own mental states. Thus, since autistic subjects do not have a properly functioning ToM they should have considerable difficulty in forming beliefs about their own mental states. So autistic people will typically not be able to form beliefs with contents like:

(1) I believe that that object is a sponge.

and

(2) I am having a visual experience of something that looks like a rock.

Perhaps (2) is too sophisticated, however. Perhaps the relevant belief that they cannot form but that normal adults can form is something more like:

(2a) That object looks like a rock.

By contrast, since ToM is *not* involved in forming beliefs about the non-mental part of the world, autistic subjects should not have great difficulty in forming beliefs like:

(3) That object is a sponge.

To get the correct answer in an appearance/reality task, subjects must have beliefs with contents like (3) and they must also have beliefs with contents like (2) or (2a). But if the TT of self-awareness is correct then autistic subjects cannot form beliefs with contents like (2) or (2a). Thus the TT predicts that autistic subjects will fail appearance/reality tasks. And since they do in fact fail, this counts as evidence in favor of the TT.

Now what we find puzzling about Argument I is that, while the data do indeed indicate that autistic subjects fail the appearance/reality task, they fail it in exactly the *wrong* way. According to Argument I, autistic subjects should have trouble forming beliefs like (2) and (2a) but should have no trouble in forming beliefs like (3). In Baron-Cohen's studies, however, just the opposite appears to be the case. After being allowed to touch and taste objects made of plastic that looked like chocolate or eggs, the autistic children gave no indication that they had incorrect beliefs about what the object *looked like*. Quite the opposite was the case. When asked questions about their own perceptual states, autistic children answered *correctly*. They reported that the fake chocolate looked like chocolate and that the fake egg looked like an egg. Where the autistic children apparently *did* have problems was just where Argument I says they should *not* have problems. The fact that they persisted in trying to eat the plastic chocolate suggests that they had not succeeded in forming beliefs like (3) -- beliefs about *what the object really is*. There are lots of hypotheses that might be explored to explain why autistic children have this problem. Perhaps autistic children have difficulty updating their beliefs on the basis of new information; perhaps they perseverate on first impressions; [10] perhaps they privilege visual information over the information provided by touch and taste; perhaps the task demands are too heavy. But whatever the explanation turns out to be, it is hard to see how the sorts of failures predicted by the TT of self-awareness -- the inability to form representations like (1), (2) and (2a) -- could have any role to play in explaining the pattern of behavior that Baron-Cohen reports.

All this may be a bit clearer if we contrast the performance of autistic children on appearance/reality tasks with the performance of normal 3 year olds. The three year olds also fail the task. But unlike the autistic children who make what Baron-Cohen calls "phenomenist" errors (Baron-Cohen 1989, 594), normal three year olds make what might

[10]It's worth noting that perseveration is quite common in autistic children in other domains as well.

be called "realist" errors on the same sorts of tasks. Once they discover that the Hollywood rock really is a sponge, they report that it *looks like* a sponge. Since there is reason to believe that ToM is not yet fully on line in 3 year olds, one might think that the fact that 3 years olds make "realist" errors in appearance/reality tasks supports a TT account of self-awareness. Indeed, Alison Gopnik appears to defend just such a view. The appearance/reality task, she argues,

> is another case in which children make errors about their current mental states as a result of their lack of a representational theory [i.e. a mature ToM] …. Although it is not usually phrased with reference to the child's current mental states, this question depends on the child's accurately reporting an aspect of his current state, namely, the way the object looks to him. Children report that the sponge-rock looks to them like a sponge. To us, the fact that the sponge looks like a rock is a part of our immediate phenomenology, not something we infer…. The inability to understand the idea of false representations… seems to keep the child from accurately reporting perceptual appearances, even though those appearances are current mental states (Gopnik 1993, 93).

For our current purposes, the crucial point here is that Gopnik's argument, unlike Argument I, is perfectly sensible. Three year olds are not at all inclined to make "phenomenist" errors on these tasks. Once they have examined the plastic chocolate, they no longer believe that it really is chocolate, and they have no inclination to eat it. Where the three year olds go wrong is in reporting what plastic chocolate and Hollywood rocks look like. And this is just what we should expect if, as the TT insists, ToM is involved in forming beliefs about one's own perceptual states.

At this point, the reader may be thinking that we have jumped out of the frying pan and into the fire. In using Gopnik's argument to highlight the shortcomings of Argument I, have we not also provided a new argument for TT, albeit one that does not rely on data about autistic subjects? Our answer here is that Gopnik's argument is certainly one that must be taken seriously. But her explanation is not the only that might be offered to account for the way in which 3 year olds behave in appearance/reality tasks. The hypothesis we favor is that though the Percept Monitoring Mechanisms that we posited in Section 3.2 are in place in 3 year olds, the usual appearance/reality tasks sidestep the PMMs because young children rely on a set of heuristic principles about appearance and reality, principles like:

> HP: For middle-sized objects, if an object is an X, then it looks like an X.

If a 3 year old relies on such a heuristic, then her answers to the questions posed in appearance/reality tasks may not engage the PMMs at all. Of course, if this defense of our PMM hypothesis is right, then there should be other ways of showing that the young children really can detect their own perceptual states. One way to explore this is by running experiments on children's understanding of appearances, but with tasks that would not implicate the above heuristic. To do this, one might ask the child questions about the relations between her perceptual mental states themselves. So, for instance, one might let children play with two different Hollywood rocks and a plain yellow sponge,

then ask the children which two look most alike. It would also be of interest to run a similar experiment based more closely on the Flavell task: Show children a plain yellow sponge, a Hollywood rock, and a piece of granite, then ask them which two look most alike. If children do well on these sorts of tasks, that would provide evidence that they can indeed detect their own perceptual states, and hence that the PMM is intact and functioning even though ToM is not yet on line.

Let us briefly sum up this section. Our major conclusion is that, while the data about the performance of autistic subjects on appearance/reality tasks are fascinating, they provide no evidence at all for the Theory Theory of self-awareness. Moreover, while the data about the performance of normal three year olds on appearance/reality tasks is compatible with the TT, there is an alternative hypotheses that is equally compatible with the data, and there are some relatively straightforward experiments that might determine which is correct.

4.2. Introspective reports and autobiographies from adults with Asperger's Syndrome

The next two arguments we will consider are much more direct arguments for the Theory Theory, but, we maintain, no more convincing. Carruthers (1996) and Frith & Happé (1999) both cite evidence from a recent study on introspective reports in adults with Asperger Syndrome (Hurlburt, Happé & Frith 1994). The study is based on a technique for "experience sampling" developed by Russell Hurlburt. Subjects carry around a beeper and are told, "Your task when you hear a beep is to attempt to 'freeze' your current experience 'in mind,' and then to write a description of that experience in a ... notebook which you will be carrying. The experience that you are to describe is the one that was occurring at the instant the beep began..." (Hurlburt 1990, 21).

Hurlburt and his colleagues had 3 Asperger adults carry out this experience sampling procedure (Hurlburt et al. 1994). All three of the subjects were able to succeed at simple theory of mind tasks. The researchers found that the reports of these subjects were considerably different from reports of normal subjects. According to Hurlburt and colleagues, two of the subjects reported only visual images, whereas it's common for normal subjects to also report inner verbalization, "unsymbolized thinking",[11] and emotional feelings. The third subject didn't report any inner experience at all in response to the beeps.

[11]Hurlburt and colleagues describe "unsymbolized thoughts" as

> clearly-apprehended, differentiated thoughts that occurred with no experience of words, images, or others symbols that might carry the meaning. Subjects sometimes referred to the phenomenon as 'pure thought'. In such samples the subjects could, in their sampling interviews, say clearly what they had been thinking about at the moment of the beep, and thus could put the thought into words, but insisted that neither those words nor any other words or symbols were available to awareness at the moment of the beep, even though the thought itself was easily apprehended at that moment (Hurlburt, Happé, & Frith 1994, 386).

Carruthers maintains that these data suggest "that autistic people might have severe difficulties of access to their own occurrent thought processes and emotions" (1996, 261). Frith and Happé also argue that the evidence "strengthens our hypothesis that self-awareness, like other awareness, is dependent on ToM." (Frith and Happé 1999, 14)

As further support for the TT account of self-awareness, Frith & Happé appeal to several autobiographical essays written by adults with autism or Asperger syndrome (1999). They argue that these autobiographies indicate that their authors have significant peculiarities in self-consciousness. Here are several examples of autobiographical excerpts quoted by Frith & Happé:

> "When I was very young I can remember that speech seemed to be of no more significance than any other sound. ... I began to understand a few single words by their appearance on paper... (Jolliffe, Lansdown & Robinson 1992, 13, quoted in Frith & Happé 1999, 15).
>
> "I had – and always had had, as long as I could remember – a great fear of jewellery... I thought they were frightening, detestable, revolting" (Gerland 1997, 54, quoted in Frith & Happé 1999, 16)
>
> "It confused me totally when someone said that he or she had seen something I had been doing in a different room" (Gerland 1997, 64, quoted in Frith & Happé 1999, 17).

4.3. What conclusions can we draw from the data on introspection in autism?

We are inclined to think that the data cited by Carruthers (1996) and Frith & Happé (1999) provide a novel and valuable perspective on the inner life of people with autism. However, we do not think that the evidence lends any support at all to the TT account of self-awareness over the MM theory that we advocate. Quite to the contrary, we are inclined to think that if the evidence favors either theory, it favors ours.

What the data do strongly suggest is that the inner lives of autistic individuals differ radically from the inner lives of most of us. Images abound, inner speech is much less salient, and autistic individuals almost certainly devote much less time to thinking or wondering or worrying about *other people's* inner lives. As we read the evidence, however, it indicates that people with autism and Asperger's Syndrome *do* have access to their inner lives. They are aware of, report and remember their own beliefs and desires as well as their occurrent thoughts and emotions.

4.3.1. Hurlburt, Happé and Frith (1994) revisited

In the experience sampling study, there were a number of instances in which subjects clearly did report their occurrent thoughts. For example, one of the subjects, Robert, reported that

> he was 'thinking about' what he had to do today. This 'thinking about' involved a series of images of the tasks he had set for himself. At the moment of the beep, he was trying to figure out how to find his way to the Cognitive Development Unit, where he had his appointment with us. This 'trying to figure out' was an image of himself walking down the street near Euston station (388).

On another occasion, Robert reported that he was

> 'trying to figure out' why a key that he had recently had made did not work. This figuring-out involved picturing an image of the key in the door lock, with his left hand holding and turning the key… The lock itself was seen both from the outside …and from the inside (he could see the levers inside the lock move as the blades of the key pushed them along) (388).

A second subject, Nelson, reported that

> he was 'thinking about' an old woman he had seen earlier that day. This thinking-about involved 'picturizing' (Nelson's own term for viewing an image of something) the old woman. . . . There was also a feeling of 'sympathy' for this woman, who (when he actually saw her earlier) was having difficulty crossing the street (390).

In all three of these cases it seems clear that the subjects are capable of reporting their current thinking and, in the latter case, their feelings. Though, as we suggested earlier, it may well be the case that the inner lives that these people are reporting are rather different from the inner lives of normal people.

Perhaps even more instructive is the fact that Hurlburt and his colleagues claim to have been surprised at how well the subjects did on the experience sampling task. Hurlburt, Happé & Frith write:

> While we had expected a relative inability to think and talk about inner experience, this was true for only one of the subjects, Peter, who was also the least advanced in terms of understanding mental states in the theory of mind battery (Hurlburt et al. 1994, 393).

Moreover, even Peter, although he had difficulty with the experience sampling method, could talk about his current experience. Thus Frith & Happé (1999) report that: "Although Peter was unable to tell us about his past inner experience using the beeper method, it was possible to discuss with him current ongoing inner experience during interviews" (1999, 14). So, far from showing that the TT account of self-awareness is correct, these data would seem to count *against* the TT account. For even Peter, who is likely to have had the most seriously abnormal ToM was capable of reporting his inner experiences.

It is true that all of the subjects had some trouble with the experience sampling task, and that one of them could not do it at all. But we think that this should be expected in subjects whose ToM is functioning poorly, *even if, as we maintain, the ToM plays no role in self-awareness.* Advocates of TT maintain that ToM plays a central role in detecting mental states in other people and in reasoning about mental states -- both their own and others'. And we are in agreement with both of these claims. It follows that people who have poorly functioning ToMs will find it difficult or impossible to attribute mental states to other people and will do little or no reasoning about mental states. So thoughts about mental states will not be very useful or salient to them. Given the limited role that thoughts about mental states play in the lives of people with defective ToMs, it is hardly surprising that, when asked to describe their experience, they sometimes do not report much. An analogy may help to make the point. Suppose two people are asked to look at a forest scene and report what they notice. One of the two is an expert on the birds of the region and knows a great deal about their habits and distribution. The other knows very little about birds and has little interest in them. Suppose further that there is something quite extraordinary in the forest scene; there is a bird there that is rarely seen in that sort of environment. We would expect that that bird would figure prominently in the expert's description, though it might not be mentioned at all in the novice's description. Now compared to autistic individuals, normal subjects are experts about mental states. They know a lot about them, they think a lot about them and they care a lot about them. So it is to be expected that autistic subjects – novices about mental states – will often fail to spontaneously mention their own mental states even if, like the person who knows little about birds, they can detect and report their own mental states if their attention is drawn to them by their interlocutor. Not surprisingly then, research on spontaneous language use in autism indicates that autistic children tend not to talk about certain mental states. Tager-Flusberg found that among children whose Mean Length of Utterance is 5 words, autistic children talk about desire and perception a good deal, and in any case much more than the children in a mental-age matched sample of children with Downs Syndrome. By contrast, however, the autistic children scarcely talked about "cognitive" mental states (e.g., believing, pretending, thinking), whereas the Downs kids did talk about such mental states (Tager-Flusberg 1993).[12]

[12] While Tager-Flusberg's findings are compatible with our theory they do raise an interesting further question: Why do these autistic subjects talk so much more about desire and perception than about belief and the other "cognitive" mental states? The answer, we think, is that attributing mental states to other people involves several separate mechanisms, some of which are spared in autism. The details make for a long story which we present elsewhere (Nichols & Stich, forthcoming a).

4.3.2. Autobiographies revisited

In the cases of autobiographical reflections, again, we maintain, a number of the examples cited by Frith and Happé are *prima facie* incompatible with the conclusion they are trying to establish. In the autobiographies, adults with autism or Asperger's syndrome repeatedly claim to recall their own childhood thoughts and other mental states. This is evident in the three quotes from Frith & Happé that we reproduced in section 4.2, and in this respect, the passages from Frith & Happé are not at all unusual. Here are three additional examples of autobiographical comments from adults with Asperger's syndrome:

> "I remember being able to understand everything that people said to me, but I could not speak back.... One day my mother wanted me to wear a hat when we were in the car. I logically thought to myself that the only way I could tell her that I did not want to wear the hat was to scream and throw it on the car floor" (Grandin 1984, 145).

> "When I was 5 years old I craved deep pressure and would daydream about mechanical devices which I could get into and be held by them.... As a child I wanted to feel the comfort of being held, but then I would shrink away for fear of losing control and being engulfed when people hugged me" (Grandin 1984, 151).

> "I didn't talk until I was almost five, you know. Before I started talking I noticed a lot of things, and now when I tell my mother she is amazed I remember them. I remember that the world was really scary and everything was over-stimulating" (reported in Dewey 1991, p. 204).

If these recollections are accurate, then these individuals must have been aware of their own mental states even though, at the time in question, they could not reliably attribute beliefs to other people.

5. Double dissociations and the Monitoring Mechanism Theory

We've argued that the evidence from autism does not support the Theory Theory of self-awareness over our theory. Indeed, it seems that the evidence provides support for our theory over the Theory Theory. In this section, we want to argue that the Monitoring Mechanism theory provides a natural explanation of a range of evidence on autism and certain other psychopathologies.

One important difference between our MM theory and all versions of the TT account of self-awareness is that on the MM theory there are two quite distinct mechanisms involved in mindreading. ToM is centrally involved in

(i) detecting other people's mental states

(ii) reasoning about mental states in other people, and

(iii) reasoning about one's own mental states.

MM is the mechanism that underlies

(iv) detecting one's own mental states.

On the TT account, by contrast, ToM is centrally involved in *all four* of these capacities.

Thus on our theory, though not on the TT, it is possible for the mechanism underlying (i)-(iii) to malfunction while the mechanism responsible for (iv) continues to function normally. It is also possible for the opposite pattern of breakdowns to occur. This would produce deficits in (iv) while leaving (i)-(iii) unaffected. One way to confirm our theory would be to find a "double dissociation" – cases in which subjects have an intact Monitoring Mechanism but a defective Theory of Mind and cases in which subjects have an intact ToM but a defective MM. If such cases could be found, that would provide evidence that there really are two independent mechanisms.

Do "double dissociations" of this sort occur? We propose that they do. In autism, we maintain, the ToM is seriously defective, though the MM is functioning normally. In patients exhibiting certain "first rank" symptoms of schizophrenia, by contrast, MM is malfunctioning though ToM is functioning normally.

5.1. Intact MM but damaged ToM

As we have already noted, it is widely agreed that autistic people have a serious ToM deficit. They fail a number of mindreading tasks, including the false belief task (see, e.g., Baron-Cohen 1995). And there is evidence that autistic individuals find minds and mental states puzzling. This comes out vividly in a famous passage from Oliver Sacks, discussing Temple Grandin:

> She was bewildered, she said, by *Romeo and Juliet* ("I never knew what they were up to"), and with *Hamlet* she got lost with the back-and-forth of the play. Though she ascribed these problems to 'sequencing difficulties,' they seemed to arise from her failure to empathize with the characters, to follow the intricate play of motive and intention. She said that she could understand "simple, strong, universal" emotions but was stumped by more complex emotions and the games people play. "Much of the time," she said, "I feel like an anthropologist on Mars" (Sacks 1995, 259).

Grandin herself writes: "My emotions are simpler than those of most people. I don't know what complex emotion in a human relationship is. I only understand simple emotions, such as fear, anger, happiness, and sadness" (Grandin 1995, 89). Similarly, although autistic children seem to understand simple desires, they have significantly more difficulty than mental-aged peers on tasks that require inferring implicit desires (Phillips et al. 1995). For instance, Phillips and her colleagues showed two comic strips to autistic children and controls. In one strip, the child is first pictured shown standing next to the pool, in the next frame the child is jumping into the pool, and in the 3rd frame the child is standing on the deck dripping wet. The other strip is basically the same as the first strip,

except that in the second frame, the child is shown falling into the pool. The children are asked, "Which boy meant to get wet?" (Phillips et al. 1995, 157). The researchers found that autistic children did considerably worse than mentally handicapped children on this task (Phillips et al. 1995). These observations and findings might reflect a difficulty in autistic individuals' ability to reason about mental states, but in any case, they certainly indicate a serious deficiency in Theory of Mind.[13] The standard view is that, unlike normal children, autistic children lack a Theory of Mind (Baron-Cohen 1995). We are skeptical that the evidence show anything quite this strong. Nonetheless, autism seems to involve a profound deficit in Theory of Mind.

Although there is abundant evidence that autism involves a deficit in Theory of Mind abilities, none of the evidence reviewed in section 4 suggests that autism involves a deficit in the Monitoring Mechanism. Indeed, some of the data suggested just the opposite. The adults with Asperger's syndrome who were asked to recount their immediate experiences did show an appreciation of what was happening in their minds (Hurlburt et al. 1994). Further, in the autobiographical excerpts, the adults claim to recall their own beliefs and thoughts from childhood. Also, there is no evidence that autistic children or adults have any trouble recognizing their thoughts and actions as their own. (The importance of this point will emerge below.) Thus, in autism, while we have good reason to think that there is a deficit in Theory of Mind, we have no reason to think that there is a deficit in the Monitoring Mechanism. All of these facts, we maintain, indicate that while autistic people suffer a serious ToM deficit, they have no deficit in the Monitoring Mechanism posited by our theory.

5.2. Intact ToM but damaged MM

We have suggested that in autism, we find a deficit in Theory of Mind but an intact Monitoring Mechanism. Are there cases in which we find the opposite pattern? That is, are there individuals with an intact Theory of Mind but a deficit in the Monitoring Mechanism? Although the data are often fragmentary and difficult to interpret, we think there might actually be such cases. Schizophrenia has recently played an important role in discussion of Theory of Mind, and we think that certain kinds of

---

[13]Since the detection/reasoning distinction hasn't been explicitly drawn in the literature, it's sometimes hard to tell whether the tasks show a deficit in detection or a deficit in reasoning. For instance, there are two standard ways to ask the false belief task question in the Maxi-task, one that is explicitly about detection (e.g., where does Maxi think the candy is?) and one that involves reasoning (e.g., where will Maxi look for the candy?). If one can't solve the detection task, then that alone would preclude one from solving the reasoning task in the right way, since one would have to detect the mental states in order to reason to the right conclusion. So, for many of the results that seem to show a deficit in reasoning about mental states, it's not yet clear whether the deficit can simply be explained by a deficit in detection. This raises a number of important issues, but fortunately, for our purposes we don't need to sort them out here.

schizophrenia might involve a damaged Monitoring Mechanism but intact Theory of Mind.

There is a cluster of symptoms in some cases of schizophrenia sometimes referred to as "passivity experiences" or "first rank symptoms" (Schneider 1959) "in which a patient's own feelings, wishes or acts seem to be alien and under external control" (C.D. Frith 1992, 73-74).

One "first rank" symptom of schizophrenia is delusions of control, in which a patient has difficulty recognizing that certain actions are her own actions. For example, one patient reported:

> "When I reach my hand for the comb it is my hand and arm which move, and my fingers pick up the pen, but I don't control them…. I sit there watching them move, and they are quite independent, what they do is nothing to do with me….I am just a puppet that is manipulated by cosmic strings. When the strings are pulled my body moves and I cannot prevent it" (Mellor 1970, 18).

Another first rank symptom is "thought withdrawal", the impression that one's thoughts are extracted from one's mind. One subject reported: "I am thinking about my mother, and suddenly my thoughts are sucked out of my mind by a phrenological vacuum extractor, and there is nothing in my mind, it is empty" (Mellor 1970, 16-17).

At least some symptomatic schizophrenics have great difficulty in reporting their current thoughts. Russell Hurlburt had four schizophrenic patients participate in a study using Hurlburt's experience sampling method (see section 4.2). Two of these subjects reported experiences and thoughts that were strange or "goofed up". One of the patients, who was symptomatic throughout the sampling period (and whose symptoms apparently included first rank symptoms), seemed incapable of carrying out the task at all. Another patient was able to carry out the task until he became symptomatic, at which point he could no longer carry out the task. Hurlburt argues that these two subjects, while they were symptomatic, did not have access to their inner experience (Hurlburt 1990, 239). Hurlburt writes:

> What we had expected to find, with Joe, was that his inner experiences were unusual – perhaps with images that were 'goofed up' as Jennifer had described, or several voices that spoke at once so that none was intelligible, or some other kind of aberrant inner experience that would explain his pressure of speech and delusions. What we found, however, was no such thing; instead, Joe could not describe *any* aspects of his inner experience in ways that we found compelling (Hurlburt 1990, 207-8).

What's especially striking here is the contrast between this claim and Hurlburt et al.'s claim about the adults with Asperger syndrome. Hurlburt (1990) expected the symptomatic schizophrenics to be able to report their inner experiences, and Hurlburt et al. (1994) expected the adults with Asperger syndrome to be unable to report their inner experiences. What they found, however, was just the opposite. The symptomatic

schizophrenics could not report their inner experiences, and the adults with Asperger syndrome could (see section 4).

These findings on schizophrenia led Christopher Frith to suggest that in schizophrenics with first rank symptoms, there is a deficit in "central monitoring" (e.g. C.D. Frith 1992, 81-82).[14] Frith's initial account of central monitoring doesn't specify how the monitoring works, but in recent work, Frith has sharpened his characterization by connecting the idea of central monitoring with the work on Theory of Mind. He suggests that the way to fill out his proposal on central monitoring is in terms of Theory of Mind.

> Many of the signs and symptoms of schizophrenia can be understood as arising from impairments in processes underlying 'theory of mind' such as the ability to represent beliefs and intentions (Frith 1994, 148).
>
> To have a "theory of mind", we must be able to represent propositions like "Chris believes that 'It is raining'". Leslie (1987) has proposed that a major requirement for such representations is a mechanism that decouples the content of the proposition (It is raining) from reality… I propose that, in certain cases of schizophrenia, something goes wrong with this decoupling process…. Failure of this decoupling mechanism would give rise … to… the serious consequence… that the patient would no longer be able to represent mental states, *either their own or those of others*. I have suggested previously (Frith 1987) that patients

---

[14]Indeed, Frith put his suggestion to the empirical test, using a series of error correction experiments to test the hypothesis that passivity experiences result from a deficit in central monitoring. Frith and colleagues designed simple video games in which subjects had to use a joystick to follow a target around a computer screen. The games were designed so that subjects would make errors, and the researchers were interested in the subjects' ability to correct the errors without external (visual) feedback indicating the error. Normal people are able to rapidly correct these errors even when they don't get feedback. Frith takes this to indicate that normal people can monitor their intended response, so that they don't need to wait for the external feedback. Thus, he suggests, "If certain patients cannot monitor their own intentions, then they should be unable to make these rapid error corrections" (C.D. Frith 1992, 83). Frith and others carried out studies of the performance of schizophrenics on these video game tasks. The researchers found that

> acute schizophrenic patients corrected their errors exactly like normal people when visual feedback was supplied but, unlike normal people often failed to correct errors when there was no feedback. Of particular interest was the observation that this disability was restricted to the patients with passivity experiences: delusions of control, thought insertion and thought blocking. These are precisely the symptoms that can most readily be explained in terms of a defect of self-monitoring (Frith 1992, 83).

Mlakar et al. (1994) found similar results. Thus, there seems to be some evidence supporting Frith's general claim that passivity experiences derive from a defect in central monitoring.

> have passivity experiences (such as delusions of control and thought insertion) because of a defect in central monitoring. Central monitoring depends on our being aware of our intention to make a particular response before the response is made. In the absence of central monitoring, responses and intentions can only be assessed by peripheral feedback. For example, if we were unable to monitor our intentions with regard to speech, we would not know what we were going to say until after we had said it. I now propose that this failure of central monitoring is the consequence of an inability to represent our own mental states, including our intentions (Frith 1994, 154; emphasis added).

Hence, Frith now views the problem of central monitoring in schizophrenia as a product of a deficit in Theory of Mind (Frith 1994). Indeed, Frith characterizes schizophrenia as late-onset autism (1994, 150).

Although we are intrigued by Frith's initial suggestion that passivity experiences derive from a deficit in central monitoring, we are quite skeptical of his claim that the root problem is a deficit in Theory of Mind. We think that a better way to fill out Frith's hypothesis is in terms of the Monitoring Mechanism. That is, we suggest that certain first rank symptoms or passivity experiences might result from a deficit in the Monitoring Mechanism that is quite independent of any deficit in Theory of Mind. And, indeed, Frith's subsequent empirical work on schizophrenia and Theory of Mind indicate that schizophrenics with passivity experiences do *not* have any special difficulty with standard theory of mind tasks. Frith & Corcoran (1996) write, "It is striking that the patients with passivity features (delusions of control, thought insertion, etc.) could answer the theory of mind questions quite well. This was also found by Corcoran et al. (1995) who used a different kind of task" (Frith & Corcoran 1996, 527). Of course, this is exactly what would be predicted by our theory since we maintain that the mechanism for detecting one's own intentions is independent from the mechanism responsible for detecting the intentions of others. Hence, there's no reason to think that a deficit in detecting one's own intentions would be correlated with a deficit in detecting intentions in others.

We maintain that, as with autism, our theory captures this range of data on schizophrenia comfortably. Contra Frith's proposal, schizophrenia does not seem to be a case in which ToM is damaged; rather, it's more plausible to suppose that in schizophrenia, it's the theory-independent Monitoring Mechanism that is not working properly. So, it's plausible that there are cases of double dissociation. Autism plausibly involves a damaged Theory of Mind without a damaged Monitoring Mechanism. And schizophrenic subjects with first rank symptoms may have a damaged Monitoring Mechanism but don't seem to have a damaged Theory of Mind. And this, we think, provides yet another reason to prefer the MM theory to the TT account of self-awareness.[15]

---

[15] The idea that a monitoring mechanism can be selectively damaged while the analogous Theory of Mind ability is intact might apply to mental states other than propositional attitudes like beliefs and desires. For instance, alexithymia is a clinical condition in which

6. The ascent routine theory

Although the Theory Theory is the most widely accepted account of self-awareness in the recent literature, there are two other accounts that are also quite visible, though neither seems to have gained many advocates. In this section and the next we will briefly consider each of these accounts.

Our MM account appeals to an innate cognitive mechanism (or a cluster of mechanisms) specialized for detecting one's own mental states. One might want to provide an account of self-awareness that is more austere. One familiar suggestion is that when we're asked a question about our own beliefs "Do you believe that p?" we treat the question as the simple fact-question: "p?". This kind of account was proposed by Evans (1982), but in recent years it has been defended most vigorously by Robert Gordon. He labels the move from belief-question to fact-question an "ascent routine" and even tries to extend the account to self-attributions of pain (Gordon 1995, 1996). Gordon writes: "self-ascription relies … on what I call *ascent routines*. For example, the way in which adults ordinarily determine whether or not they believe that p is simply to ask themselves the question whether or not p." (Gordon 1996, 15).

This account has the virtue of emphasizing that, for both children and adults, questions like "Do you think that p?" and "Do you believe that p?" may not be interpreted as questions about one's mental state, but as questions about p. Similarly, statements like "I believe that p" are often guarded assertions of p, rather than assertions about the speaker's mental state.[16] These are facts that must be kept in mind in interpreting the results of experiments on mindreading and self-awareness.

---

subjects have great difficulty discerning their own emotional states. One researcher characterizes the condition as follows: "When asked about feelings related to highly charged emotional events, such as the loss of a job or the death of a family member, patients with alexithymia usually respond in one of two ways: either they describe their physical symptoms or they seem not to understand the question" (Lesser 1985, 690). As a result, patients with this condition often need to be given instruction about how to interpret their own somatic sensations. "For instance, they need to understand that when one is upset or scared, it is normal to feel abdominal discomfort or a rapid heart beat. These sensations can be labeled 'anger' or 'fear'" (691). Thus alexithymia might be a case in which subjects have selective damage to a system for monitoring one's own emotions. Of course, to make a persuasive case for this, one would need to explore (among other things) these subjects' ability to attribute emotions to other people. If it turns out that patients with alexithymia can effectively attribute emotions to others but not to themselves, that would indicate that alexithymia might indeed be caused by damage to a monitoring mechanism. We think that these kinds of questions and experiments only become salient when we get clear about the distinction between theory of mind and monitoring mechanisms. (We are grateful for Robert Woolfolk suggesting this interpretation of alexithymia.)

[16] Claims like this are, of course, commonplace in the philosophical literature on the "analysis" of belief. For example, Urmson maintains that 'believe' is a 'parenthetical

Alongside these virtues, however, the ascent routine also has clear, and we think fatal, shortcomings. As Goldman (forthcoming) points out, the ascent routine story doesn't work well for attitudes other than belief.

> Suppose someone is asked the question, 'Do you hope that Team T won their game yesterday?' (Q_1). How is she supposed to answer that question using an ascent routine? Clearly she is not supposed to ask herself the question, "Did Team T win their game yesterday?" (Q_2), which would only be relevant to belief, not hope. What question is she supposed to ask herself? (Goldman forthcoming, 23).

Furthermore, even for beliefs and thoughts, the ascent routine strategy doesn't work for some central cases. In addition to questions like "do you believe that p?", we can answer questions about current mental states like "what are you thinking about?". But in this case, it is hard to see how to rework the question into an ascent routine. Similarly, as we noted earlier, people can give accurate retrospective reports in response to questions like "how did you figure that out?" We can see no way of transforming these questions into fact-questions of the sort that Gordon's theory requires. This also holds for questions about current desires, intentions, and imaginings, questions like: "what do you want to do?"; "what are you going to do?"; "what are you imagining?" Our ability to answer these questions suggests that the ascent routine strategy simply can't accommodate many central cases of self-awareness. There is no plausible way of recasting these questions so that they are questions about the world rather than about one's mental state. As a result, the ascent routine account strikes us as clearly inadequate as a general theory of self-awareness.

7. The phenomenological theory

For the last several years, Alvin Goldman has been advocating a "phenomenological model for the attitudes" (Goldman 1993b, 23; see also Goldman 1997, forthcoming). According to Goldman, in order to detect one's own mental states, "the cognitive system [must] use … information about the *intrinsic* (nonrelational) and *categorical* (nondispositional) properties of the target state" (1993a, 87) Goldman then goes on to ask "which intrinsic and categorical properties might be detected in the case of mental states?" His answer is as follows: "The best candidates, it would seem, are so-called *qualitative properties* of mental states – their phenomenological or subjective

---

verb' and that such verbs "*are not psychological descriptions*" (Urmson 1956, 194). Rather, "when a man says, 'I believe that he is at home' or 'He is, I believe, at home', he both implies a (guarded) claim of the truth, and also implies a claim of the reasonableness of the statement that he is at home" (Urmson 1956, 202).

feelings (often called "qualia")". (1993a, 87)[17] So, on this view, one detects one's own mental states by discerning the phenomenological properties of the mental states, the way those mental states feel.

Goldman is most confident of this phenomenological approach when the mental states being detected are not propositional attitudes but rather what he calls "sensations." "Certainly," he argues, "it is highly plausible that one classifies such sensations as headaches or itches on the basis of their qualitative feel" (1993a, 87). Goldman suggests that this account might also be extended to propositional attitudes, though he is rather more tentative about this application.

> Whether the qualitative or phenomenological approach to mental concepts could be extended from sensations to attitudes is an open question. Even this prospect, though, is not beyond the bounds of credibility. There is no reason why phenomenological characteristics should be restricted to sensory characteristics, and it does indeed seem to "feel" a particular way to experience doubt, surprise, or disappointment, all of which are forms of propositional attitudes. (1993a, 88; see also 1993b, 25, 104).

We are inclined to think that the idea of extending the phenomenological approach from sensations to propositional attitudes is much less of an "open question" than Goldman suggests. Indeed, as a general theory of the self-attribution of propositional attitudes, we think that it is quite hopeless.

7.1. Two versions of Goldman's proposal

Let us begin by noting that there are two quite different ways in which Goldman's proposal might be elaborated:

> a. *The Weaker Version* claims that we (or our cognitive systems) detect or classify the *type* of a given mental state by the qualitative or phenomenological properties of the mental state in question. It is the qualitative character of a state that tells us that it is a belief or a desire or a doubt. On the weaker version, however, the qualitative properties of propositional attitudes do not play a role in detecting the *content* of propositional attitudes.
>
> b. *The Stronger Version* claims that we (or our cognitive systems) detect or classify *both* the *type* and *the content* of a given mental state by the qualitative or phenomenological properties of the mental state in question. So it is the qualitative character of a state that tells us that it is a belief or a desire and it is also the qualitative character that tells us that it is the belief *that there is no greatest prime number* or the desire that the Democrats win the next election.

[17] Since Goldman regards these phenomenological properties as 'intrinsic', he rejects the higher-order account of consciousness advocated by Rosenthal (1992) and others (see Goldman, forthcoming, 17).

If one speaks, as we just did, of qualitative or phenomenological qualities "telling us" that a state is a belief or that its content is *that there is no greatest prime number*, it is easy to ignore the fact that this is a metaphor. Qualitative states don't literally "tell" anybody anything. What is really needed, to make a proposal like Goldman's work, is a mental mechanism (or a pair of mental mechanisms) which can be thought of as transducers: They are acted upon by the qualitative properties in question and produce, as output, *representations* of these qualitative properties (or, perhaps more accurately, representations of the kind of state that *has* the qualitative property). So, for example, on the Weaker Version of the theory, what is needed is a mechanism that goes from the qualitative property associated with belief or doubt to a representation that the state in question is a belief or doubt. On the Stronger Version, the transducer must do this for the content of the state as well. So, for instance, on the Stronger Version, the transducer must go from the qualitative property of the content *there is no greatest prime number* to a representation that the state in question has the content *there is no greatest prime number*. Figure 9 is an attempt to depict the mechanisms and processes required by Goldman's theory.

FIGURE 9 ABOUT HERE

7.2. Critique of the Goldman's theory

As we see it, the Weaker Version of Goldman's proposal is not a serious competitor for our MM theory, since the Weaker Version does not really explain some of the crucial facts about self-awareness. At best, it explains how, if I know that I have a mental state with the content *p*, I can come to know that it is a belief and not a hope or desire. But the Weaker Version doesn't even try to explain how I know that I have a mental state with the content *p* in the first place. So as a full account of self-awareness of propositional attitudes, the weaker version is a non-starter.

The Stronger Version of Goldman's model *does* attempt to provide a full account of self-awareness of propositional attitudes. However, we think that there is no reason to believe the account, and there is good reason to doubt it.

The Stronger Version of Goldman's theory requires a phenomenological account of the awareness of content as well as a phenomenological account of the awareness of attitude type. Goldman does not provide a detailed argument for a phenomenological account of content, but he does sketch one argument in its favor. The argument draws on an example proposed by Keith Gunderson (1993). Goldman discusses the example as follows:

> If I overhear Brown say to Jones, "I'm off to the bank," I may wish to know whether he means a spot for fishing or a place to do financial transactions. But if *I* say to someone, "I'm off to the bank," I cannot query my own remark: "To go fishing or to make a deposit?" I virtually always already know. . . The target article mainly supported a distinctive phenomenology for the attitude types. Gunderson's example supports distinctive phenomenology for different *contents* (Goldman 1993b, 104).

We think this argument is wholly unconvincing. It's true that we typically know the interpretation of our own ambiguous sentences. However, this doesn't even begin to show that belief contents have distinctive phenomenologies. At best it shows that we must have *some* mechanism or strategy for obtaining this knowledge. The MM theory can quite comfortably capture the fact that we typically know the interpretations of our own ambiguous sentences, and it does so without resorting to phenomenological features of content. As far as we can tell, then, there is no reason to adopt the phenomenological account of content. Moreover, there are two rather obvious reasons to prefer the MM account to the Stronger Version of the Phenomenological Theory.

On an account like Goldman's there must be mechanisms in the mind that are sensitive to phenomenological or qualitative properties – i.e. mechanisms that *are causally affected by* these qualitative properties in a highly sensitive and discriminating way. The qualia of a belief must lead the mechanism to produce a representation of belief. The qualitative properties of states with the content *Socrates is wise* must cause the mechanism to produce representations with the content *Socrates is wise*. Now we don't wish to claim that there are no mechanisms of this sort or that there couldn't be. But what is clear is that no one has a clue about how such mechanisms would work. No one has even the beginning of a serious idea about how a mechanism could be built that would be differentially sensitive to the (putative) qualitative properties of the contents of propositional attitude states. So, for the moment, at least, the mechanisms that Goldman needs are quite mysterious. The mechanism that *our* theory needs, by contrast, is simple and straightforward. To generate representations of one's own beliefs, all that the Monitoring Mechanism has to do is copy representations in the Belief Box, embed them in a representation schema of the form: *I believe that* ___., and then place this new representation back in the Belief Box. The analogous sort of transformation for representations in a computer memory could be performed by a simple and utterly *un*mysterious mechanism.[18]

The preceding argument is simply that it would be trivial to implement a mechanism like the MM whereas no one has the faintest idea how to implement the mechanisms required for Goldman's account or how such mechanisms could work. Of course, this is just a prima facie argument against Goldman's account. If it were independently plausible that phenomenology is the basis for awareness of one's own propositional attitudes, then the mysteriousness of the transducers would simply pose a challenge for cognitive scientists to figure out how such a mechanism could work. However, far from being independently plausible, it seems to us that the phenomenological account is *phenomenologically implausible.* To take the Stronger Version of Goldman's proposal seriously, one would have to assume that there is a

[18] It might be argued that the PMM that we posit in section 3.2 is just as mysterious as the mechanism that Goldman's theory requires. However, nothing in our account of the PMM requires that it is sensitive to *qualitative* properties of percepts. But even if it turns out that the PMM is sensitive to qualitative properties, we are inclined to think that the objection that we are proposing in this paragraph still has some force, since Goldman's account invokes a rather mysterious mechanism when a very unmysterious one would do the job.

distinct feel or qualia for every *type* of propositional attitude, *and* a distinct qualia for every content (or at least for every content we can detect). Now perhaps others have mental lives that are very different from ours. But from our perspective this seems to be (as Jerry Fodor might say) *crazy*. As best we can tell, believing that 17 is a prime number doesn't feel any different from believing that 19 is a prime number. Indeed, as best we can tell, neither of these states has any distinctive qualitative properties. Neither of them feels like much at all. If this is right, then the Strong Version of the Phenomenological Theory is every bit as much a non-starter as the Weak Version.

8. 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 other authors have drawn the wrong conclusions from the data. Contrary to the claims of Theory Theorists, the 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. The intriguing evidence from autism has been used to support the Theory Theory. But we've argued that the evidence from psychopathologies actually suggests the opposite. The available evidence indicates that the capacity for understanding other minds can be dissociated from the capacity to detect one's own mental states and that the dissociation can go in either direction. If this is right, it poses a serious challenge to the Theory Theory, but it fits neatly with our suggestion that the Monitoring Mechanism is independent of the Theory of Mind. Like our Monitoring Mechanism theory, the ascent routine and the phenomenological accounts are also alternatives to the Theory Theory; but these theories, we have argued, are either obviously implausible or patently insufficient to capture central cases of self-awareness. 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, Christopher Frith, Gary Gates, Rochel Gelman, Alison Gopnik, Robert Gordon, 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. (1995). *Mindblindness*. Cambridge, MA: MIT Press.

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.

Corcoran, R. Frith, C., and Mercer, G. (1995). Schizophrenia, symptomatology and social inference: Investigating 'theory of mind' in people with schizophrenia. *Schizophrenia Research*, 17, 5-13.

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

Dewey, M. (1991). Living with Asperger's syndrome. In: *Autism and Asperger Syndrome*, ed. Uta Frith. Cambridge: Cambridge University Press.

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

Evans, G. (1982). *The varieties of reference*. Oxford: Oxford University Press.

Flavell, J., Green, F., Flavell, E. (1986). *Development of knowledge about the appearance-reality distinction.* Chicago, Ill. : Society for Research in Child Development.

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

Frith, C. (1987). The positive and negative symptoms of schizophrenia reflect impairment in the perception and initiation of action. *Psychological Medicine*, 17, 631-648.

Frith, C. (1992). *The cognitive neuropsychology of schizophrenia*. Hillsdale, NJ: LEA.

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

Frith, C. & Corcoran, R. (1996). Exploring 'theory of mind' in people with schizophrenia. *Psychological Medicine*, 26, 521-530.

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

Frith, U. (1991). *Autism and Asperger syndrome*. Cambridge: Cambridge University Press.

Gerland, G. (1997). *A real person: Life on the outside*. Translated from the Swedish by J. Tate. London: Souvenir Press.

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. (forthcoming). 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. & 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. & 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.

Grandin, T. (1984). My experiences as an autistic child and review of selected literature. *Journal of Orthomolecular Psychiatry*, 13, 144-175.

Grandin, T. (1995). *Thinking in pictures*. New York: Doubleday.

Gunderson, K. (1993). On behalf of phenomenological parity for the attitudes. *Behavioral and Brain Sciences*, 16, 46-7.

Hurlburt, R. (1990). *Sampling normal and schizophrenic inner experience*. New York : Plenum Press.

Hurlburt, R., Happé, F. & Frith, U. (1994). Sampling the form of inner experience in three adults with Asperger syndrome. *Psychological Medicine*, 24, 385-395.

Jolliffe, T., Lansdown, R. & Robinson, C. (1992). Autism: A personal account. *Communication*, 26, 12-19.

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

Lesser, I. (1985). Current concepts in psychiatry. *The New England Journal of Medicine*, 312, 690-692.

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

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

Mellor, C. (1970). First rank symptoms of schizophrenia. *British Journal of Psychiatry*, 117, 15-23.

Mlakar, J, Jensterle, J. & Frith, C. (1994). Central monitoring deficiency and schizophrenic symptoms. *Psychological Medicine*, 24, 557-564.

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. (forthcoming a). *Mindreading*. Oxford: Oxford University Press.

Nichols, S. and Stich, S. (forthcoming b). Reading One's Own Mind: Self-Awareness and Developmental Psychology. In: *Working Through Thought*, ed. R. Kukla, R. Manning and R. Stainton. Boulder, CO: Westview 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. In: *Theories of Theories of Mind*, ed. 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.

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.

Phillips, W., Baron-Cohen, S. & Rutter, M. (1995). To what extent can children with autism understand desire?. *Development and Psychopathology*, 7, 151-169.

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

Rosenthal, D. (1992). Thinking that one thinks. In: *Consciousness*, ed. M. Davies and G. Humphreys. Oxford: Blackwell.

Sacks, O. (1995). *An anthropologist on mars*. New York: Alfred A. Knopf.

Schneider, K. (1959). *Clinical psychopathology*. Trans. M. Hamilton. New York: Grune & Stratton.

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.

Shallice, T. (1988). *From Neuropsychology to Mental Structure*. Cambridge: Cambridge University Press.

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: Blackwell, 87-108.

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.

Tager-Flusberg, H. (1993). What language reveals about the understanding of minds in children with autism. In: *Understanding other minds: Perspectives from autism*, ed. S. Baron-Cohen, H. Tager-Flusberg, and D. Cohen. Oxford: Oxford University Press.

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

Urmson, J. (1956). Parenthetical verbs. In: *Essays in Conceptual Analysis*, ed. A. Flew. London: MacMillan.

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.

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.