RuCCS TR-50
April, 1999

Reading One's Own Mind: A Cognitive Theory of Self Awareness

Shaun Nichols
nichols@cofc.edu
College of Charleston
Center for Cognitive Science
Rutgers University

Stephen Stich
stich@ruccs.rutgers.edu
Rutgers University
Center for Cognitive Science

Technical Report TR-50
Center for Cognitive Science
& Lab of Vision Research
Psych Bldg Addition, Busch Campus
Rutgers University - New Brunswick
152 Frelinghuysen Road
Piscataway, NJ 08854-8020

Reading One's Own Mind:
A Cognitive Theory of Self-Awareness

Shaun Nichols
(College of Charleston & Rutgers University)
nichols@cofc.edu

Stephen Stich
(Rutgers University)
stich@ruccs.rutgers.edu

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 "mind reading", the process of attributing mental states to people (and other organisms). During the last 15 years, the processes underlying mind reading 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 cognitive mechanism that plays a central role in attributing mental states to others. That mechanism 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 forthcoming), 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 serious 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 sections 4 and 5, we will review these arguments and try to show that none of the evidence favors the TT over our Monitoring Mechanism Theory. Indeed, we'll maintain that a closer look at the evidence on development and psychopathologies actually provides arguments *against* the TT. In the sixth section, we will provide a further 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 7, 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 (1993, 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.

Mind reading 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* mental states to someone.

b. *Reasoning* is the capacity to *use* information about a person's mental states to make predictions about the person's further mental states (past, present & future), 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 mind reading 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).

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 forthcoming)[1]. These issues are largely

[1]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 the different external environment.

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 "boxological" rendition of the basic architecture assumption.

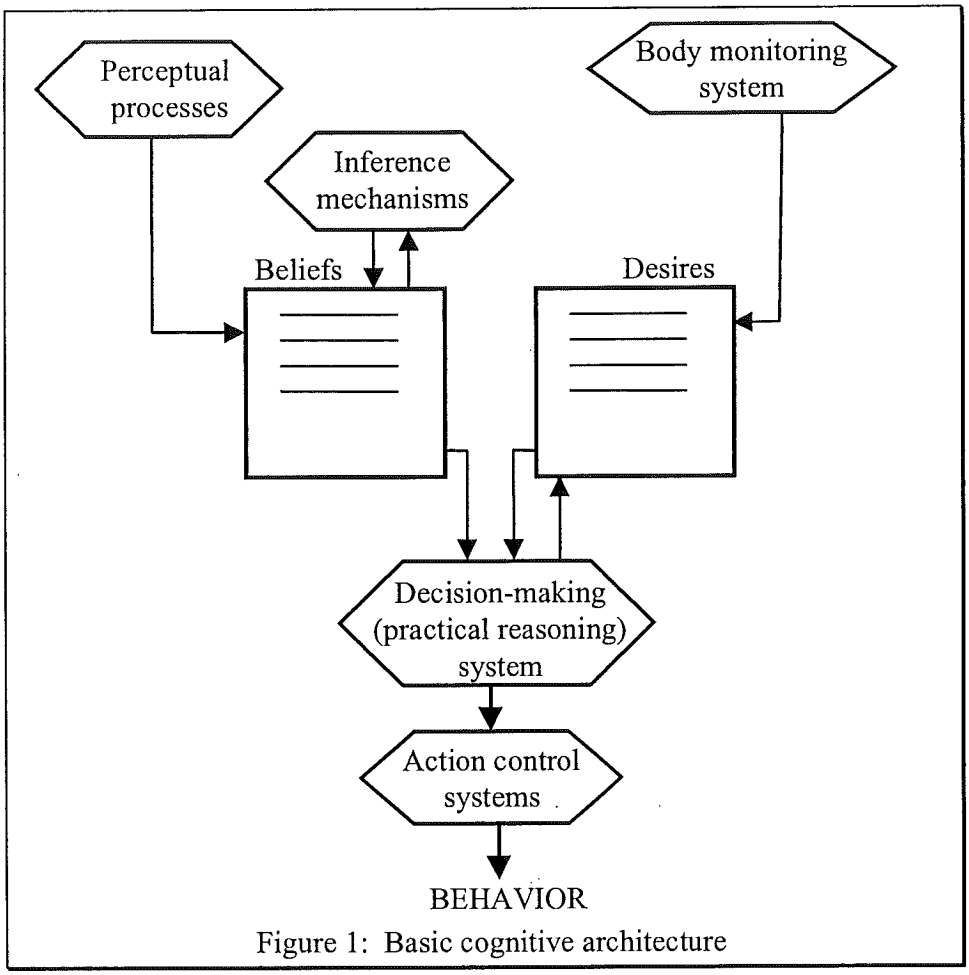

Figure 1: Basic cognitive architecture

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

Our second assumption, which we'll call *the representational account of cognition*, maintains that beliefs, desires and other propositional attitudes are relational states. To have a belief or a desire with a particular content is to have a representation token with that content stored in the functionally appropriate way in the mind. So, for example, to believe that Socrates was an Athenian is to have a representation token whose content is *Socrates was an Athenian* stored in one's Belief Box, and to desire that it will be sunny tomorrow is to have a representation whose content is *It will be sunny tomorrow* stored in one's Desire Box. Many advocates of the representational account of cognition also assume that the representation tokens

subserving propositional attitudes are linguistic or quasi-linguistic in form. This additional assumption is no part of our theory, however. If it turns out that some propositional attitudes are subserved by representation tokens that are not plausibly viewed as having a quasi-linguistic structure, that's fine with us.

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

2. The Theory Theory

As noted earlier, the prevailing account of self-awareness is the Theory Theory (TT). 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.[2]

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) 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 boxological sketch of this account is given in Figure 2.

---

[2]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 six 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 accommodate the facts about self-awareness. So, until section 7, we will simply assume that reasoning about other minds depends on a theory.

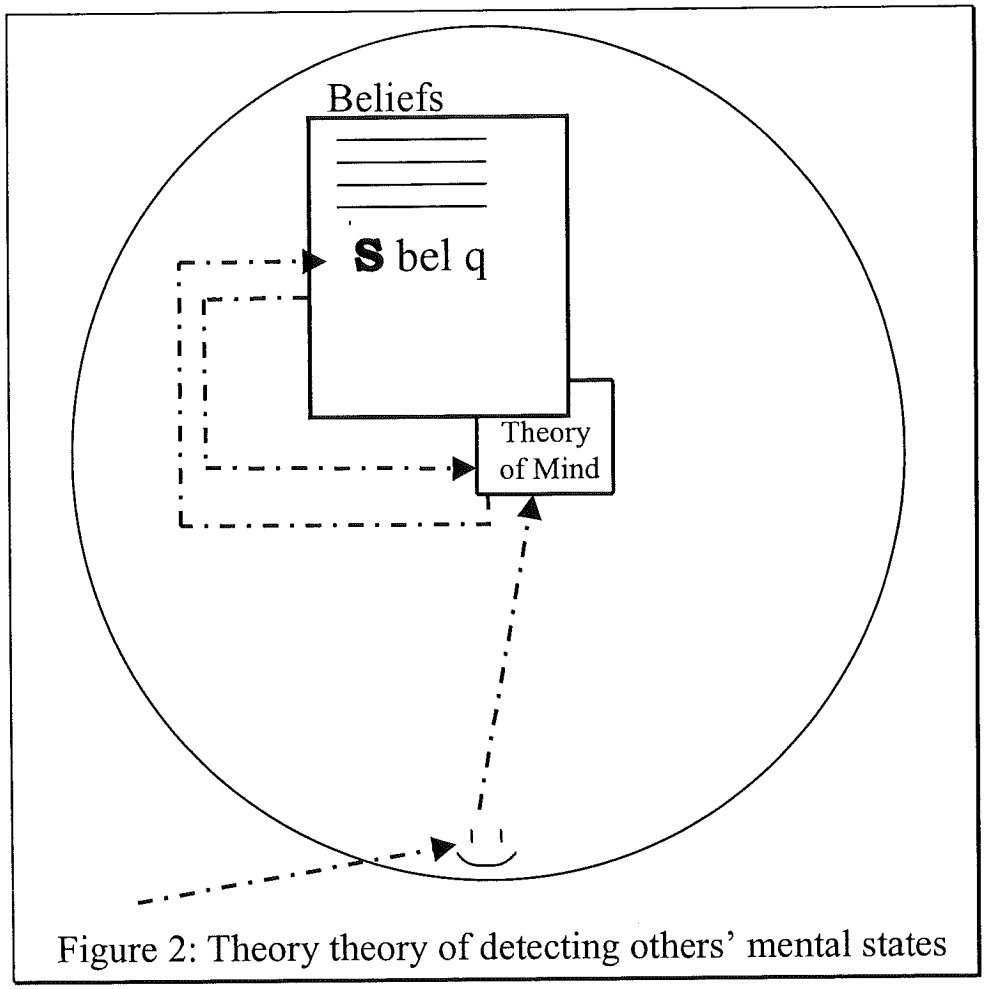

Figure 2: Theory theory of detecting others' mental states

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 have been produced by detection inferences. When detecting and reasoning are depicted together we get Figure 3.

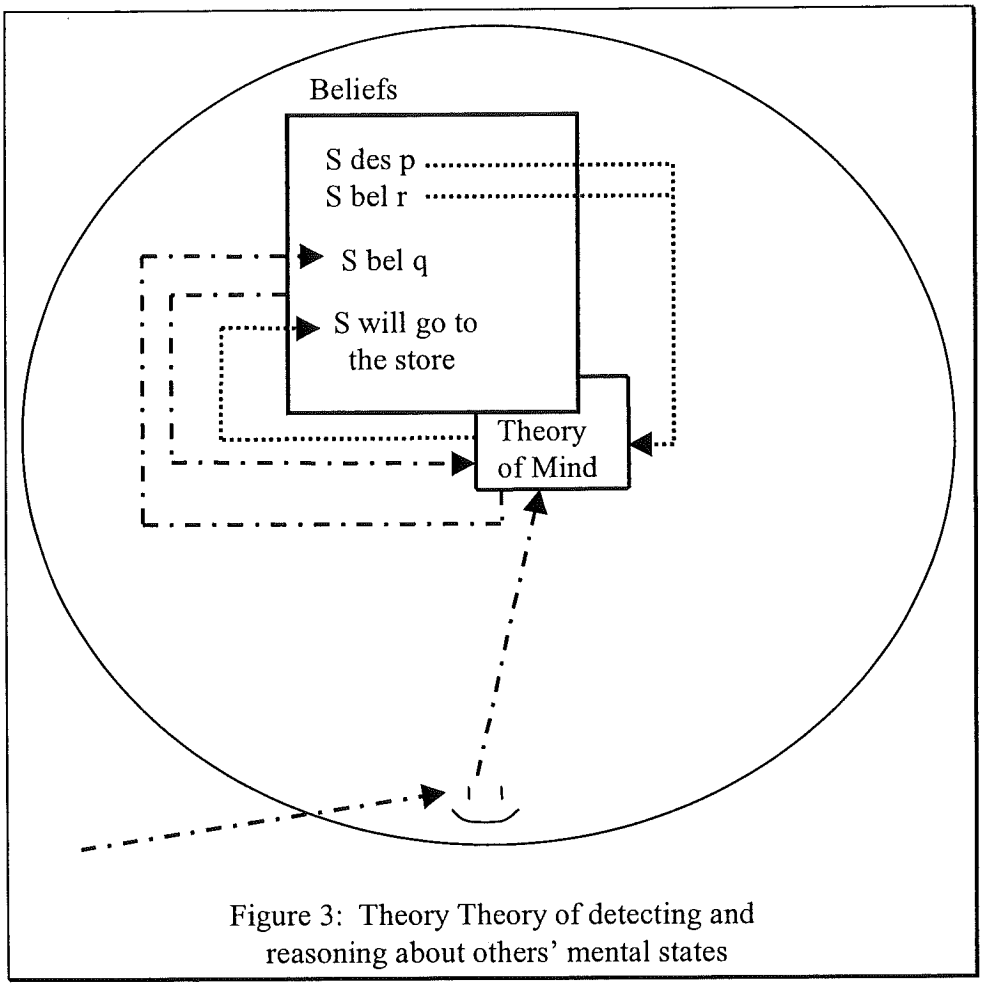

Figure 3: Theory Theory of detecting and reasoning about others' mental states

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:

> 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 forthcoming, 6).

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 with 3rd person reports 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.

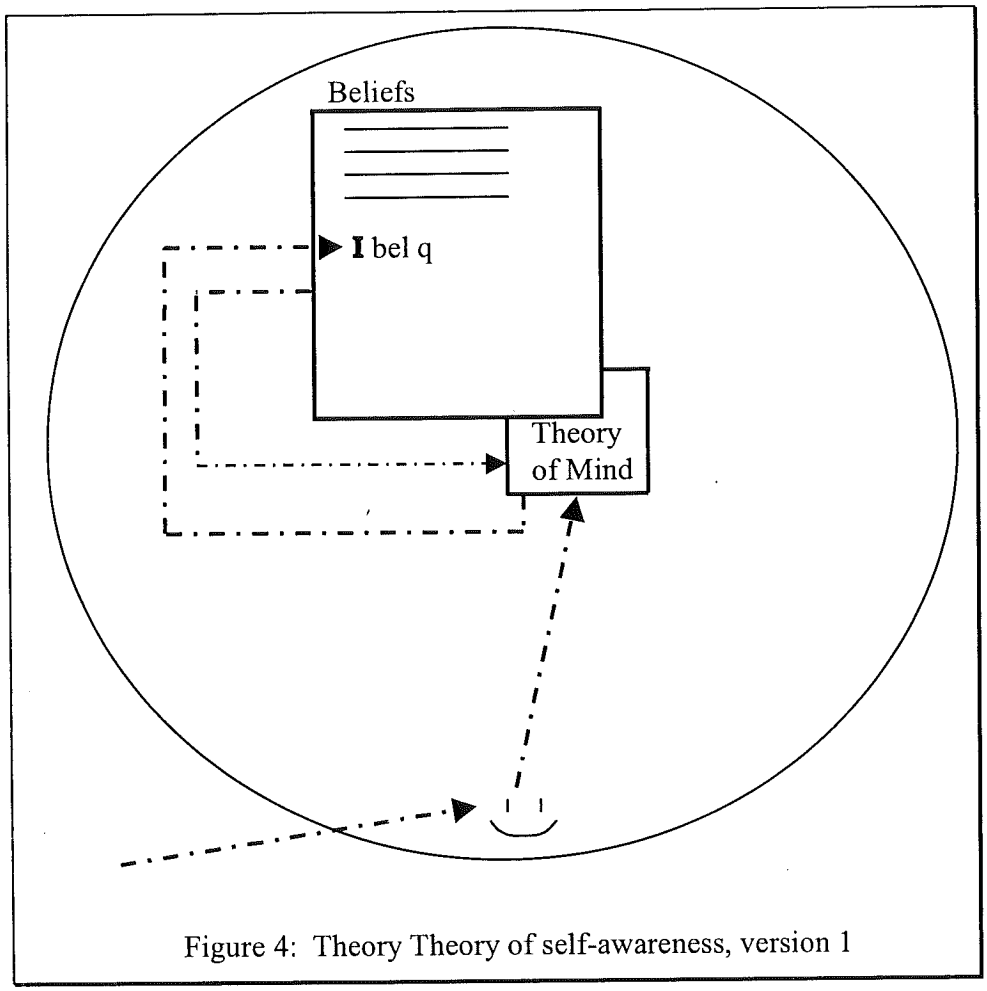

Figure 4: Theory Theory of self-awareness, version 1

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. This is at odds with the 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.[3] For instance, people are 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[4]. 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).

---

[3]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).

[4]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).

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

2.2.2. Theory Theory Version 2

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

> one possible source of evidence for the child's theory may be first-person psychological experiences that may themselves be the consequence of genuine psychological perceptions. For example, we may well be equipped to detect certain kinds of internal cognitive activity in a vague and unspecified way, what we might call *"the Cartesian buzz"* (Gopnik 1993, 11).

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

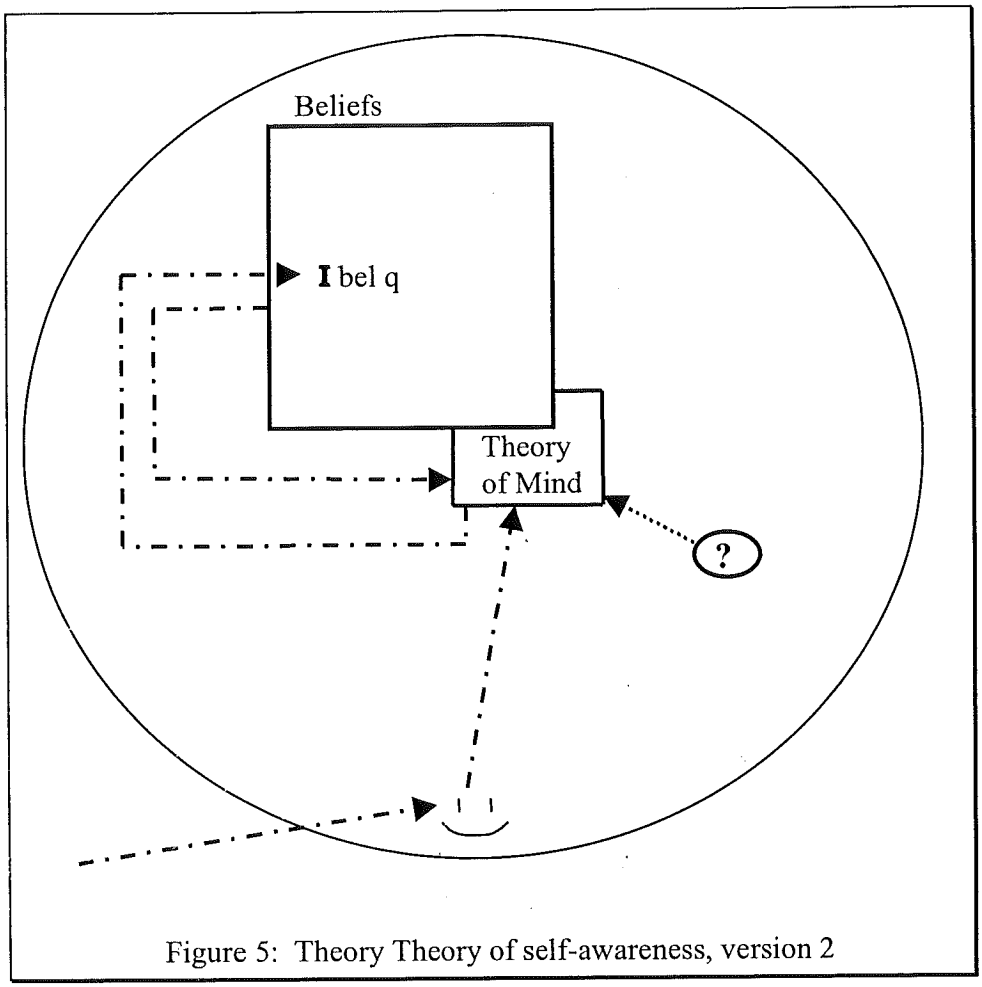

Figure 5: Theory Theory of self-awareness, version 2

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.

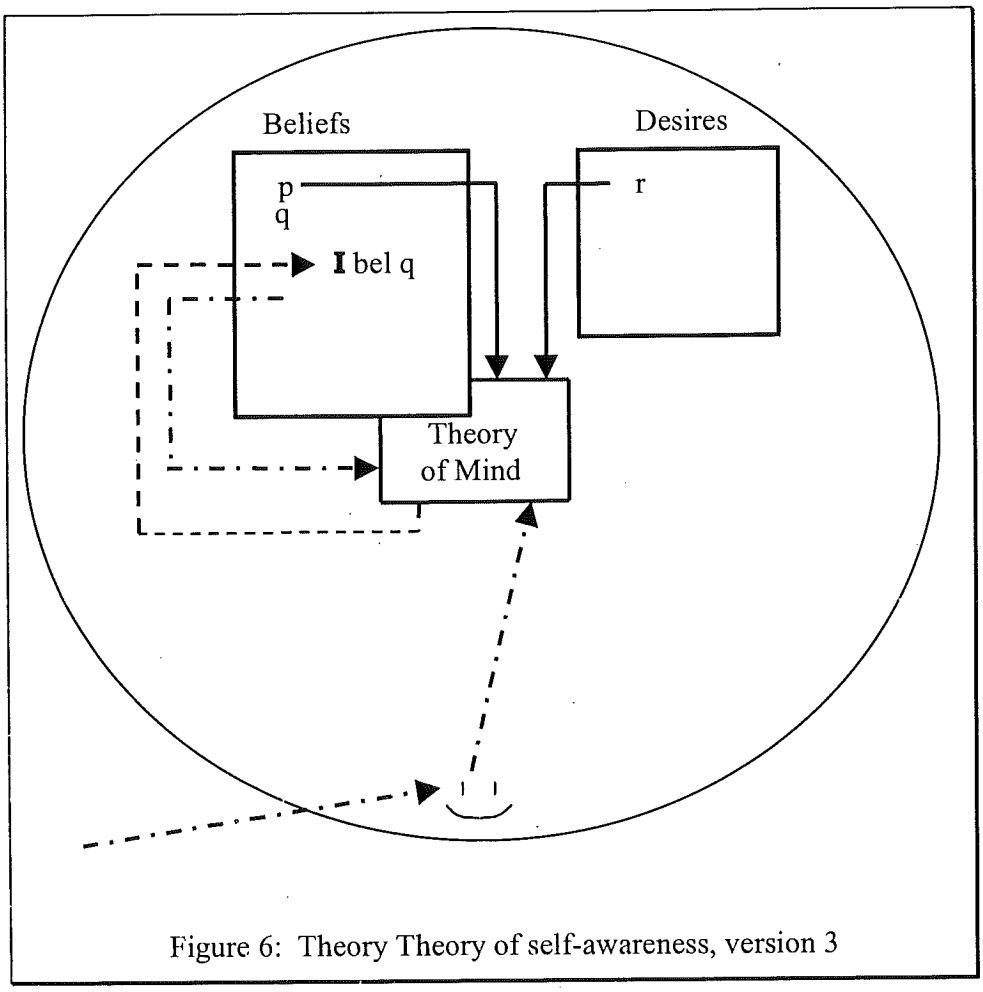

Figure 6: Theory Theory of self-awareness, version 3

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

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

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

3.1. The Monitoring Mechanism and propositional attitudes

Recall what the theory of self-awareness needs to explain. The basic facts are that when normal adults believe that p, they can quickly and accurately form the belief *I believe that p*; when normal adults desire that p, they can quickly and accurately form the belief *I desire that p*; and so on for the rest of the propositional attitudes. In order to implement this ability, no sophisticated Theory of Mind is required. All that is required is that there be a Monitoring Mechanism (MM) (or perhaps a set of mechanisms) that, when activated, takes the representation *p* in the Belief Box as input and produces the representation *I believe that p* as output. This mechanism would be trivial to implement. To produce representations of one's own beliefs, the Monitoring Mechanism merely has to copy representations from the Belief Box, embed the copies in a representation schema of the form: *I believe that* ___., and then place the new representations back in the Belief Box. The proposed mechanism would work in much the same way to produce representations of one's own desires, intentions, and imaginings.[5] This account of the process of self-awareness is sketched in Figure 7.

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

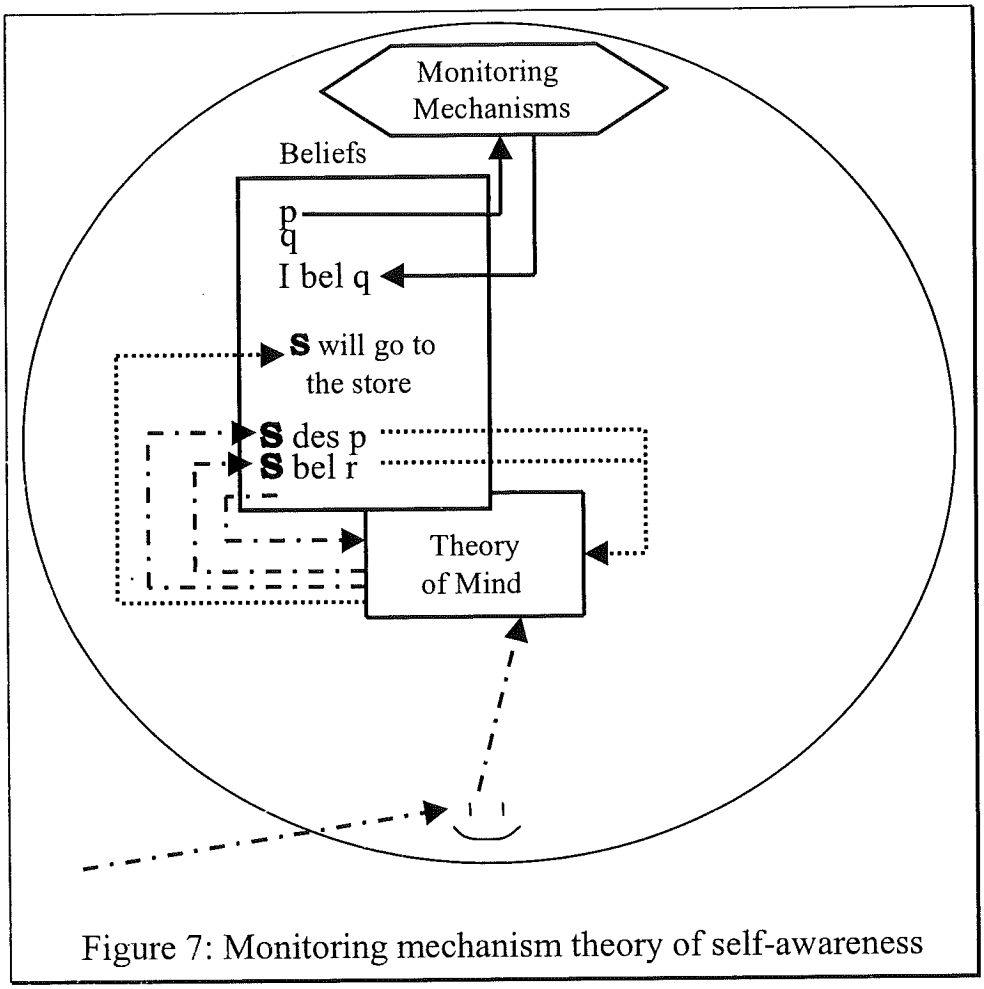

Figure 7: Monitoring mechanism theory of self-awareness

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

1977). This explanation of the findings fits well with our account too. For on our account, when trying to figure out the *causes* of one's own behavior, one must reason about mental states, and this process is mediated by the ToM. As a result, if the ToM is not up to the task, then people will make mistakes in reasoning about their own mental states as well as others' mental states.

In this paper, we propose to remain agnostic about the extent to which ToM is innate. However, we do propose that the MM (or cluster of MMs) is innate and comes on line fairly early in development – significantly before ToM is fully in place. During the period when the Monitoring Mechanism is up and running but ToM is not, the representations that the MM produces can't do much. In particular, they can't serve as premises for reasoning about mental states, since reasoning about mental states is a process mediated by ToM. So, for example, ToM provides the additional premises (or 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 this, 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 the project of constructing theories of content is deeply misguided [Stich 1992, 1996].) But, with this caveat, we don't have any objection to the claim that early *I believe that p* representations don't have the content: *I believe that p*. If that's what your favorite theory of content says, that's fine with us. Our proposal can be easily rendered consistent with such a view of content by simply replacing the embedded mental predicates (e.g., "*believe*") with technical terms "bel", "des", "pret", etc. We might then say that the MM produces the belief that *I bel that p* and the belief that *I des that q*; and that at some point further on in development, these beliefs acquire the content *I believe that p*, *I desire that q*, and so forth. That said, we propose to ignore this subtlety for the rest of the paper.

The core claim of our theory is that the MM is a distinct mechanism that is specialized for detecting one's own mental states.[6] However, it is important to note that on our account of mindreading, the MM is not the *only* mental mechanism that can generate representations with the content *I believe that p*. Representations of this sort can also be generated by ToM. Thus it

---

[6]As we've presented our theory, the MM is a mechanism that is distinct from the ToM. But it might be claimed that the MM that we postulate is just a *part* of the ToM. Here the crucial question to ask is whether it is a "dissociable" part which could be selectively damaged or selectively spared. If the answer is no, then we'll argue against this view in section 6. 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.