This article appeared in *Philosophical Topics*, 28, 171-199.

The Mind's "I" and the *Theory of Mind*'s "I": Introspection and Two Concepts of Self[1]

Shaun Nichols

College of Charleston

I. INTRODUCTION

Introspection plays a crucial role in Modern philosophy in two different ways. From the beginnings of Modern philosophy, introspection has been used a tool for philosophical exploration in a variety of thought experiments. But Modern philosophers (e.g., Locke and Hume) also tried to characterize the nature of introspection as a psychological phenomenon. In contemporary philosophy, introspection is still frequently used in thought experiments. And in the analytic tradition, philosophers have tried to characterize conceptually necessary features of introspection.[2] But over the last several decades, philosophers have devoted relatively little attention to the cognitive characteristics of introspection. This has begun to change, impelled largely by a fascinating body of work on how children and autistic individuals understand the mind.[3] In a pair of recent papers, Stephen Stich and I have drawn on this empirical work to develop an account of introspection or self-awareness.[4] In this paper, I will elaborate and defend this cognitive theory of introspection further and argue that if the account is right, it may have important ramifications for psychological and philosophical debates over the self.

Since the paper will cover a rather diverse set of issues, let me begin by mapping out the structure of what follows. In section II, I will set out the most prominent account of introspection in the recent literature, the Theory Theory of self-awareness, according to which

the capacity to detect one's own mental states depends on the capacity to detect other people's mental states. I'll then sketch the alternative "Monitoring Mechanism" account that Stich and I have defended. I will go on to offer a couple of new arguments for the Monitoring Mechanism account, and I will argue that the Monitoring Mechanism is plausibly modularized to an interesting extent. I will also respond to the worry that evolutionary considerations cast doubt on the theory. In section III, I review psychological work on the concept of self, and I argue that the Monitoring Mechanism theory suggests that there is an important notion of self that is largely neglected in the psychological literature and that needs to be distinguished from a concept of self that derives from the Theory of Mind. In the 4th section, I argue that this distinction between two concepts of self helps to explain recalcitrant philosophical problems concerning the self.

II. THE MODULARITY OF THE MIND'S EYE

The best known and most influential account of self-awareness in the recent literature is the *Theory Theory* of self awareness.[5] According to this account, one determines one's own mental states by using a "theory" of mind, and this theory of mind is the very same theory that one uses to determine the mental states of others. In philosophy, the Theory Theory of self-awareness was first proposed by Sellars,[6] but its growing influence in cognitive science is largely due to the work in developmental psychology on the understanding of other minds or *mindreading*.

To make the view clear, it is important to review a bit of the history in developmental psychology. The prevailing view of how children (and adults) understand other minds is that there is a body of information that guides psychological attribution, prediction, and explanation.

This body of information is often referred to as the child's "Theory of Mind", and it has been the subject of intense empirical and theoretical investigations. The bulk of this research explores the child's developing capacity to understand the beliefs, desires, and perceptions of others. For instance, the best known result in this area is that children under the age of 4 tend to fail the "false belief task". In one version of the false belief task, the child is shown a candy-box and asked what she thinks is in the box. After the child says that there is candy in the box, she is shown that, in fact, there are pencils in the box. The box is then closed and the child is asked what another person (who is not present) will think is in the box. Three year olds tend to say that the other person will think that there are pencils in the box, while children over the age of 4 tend to answer correctly that the other person will think that there is candy in the box. [7] These sorts of findings are taken as evidence for the development or maturation of the child's Theory of Mind.

The core idea of the Theory Theory of self-awareness is that the child's capacity for understanding her own mind depends on the same Theory of Mind that she uses to understand other minds. Alison Gopnik has been perhaps the most visible advocate for this view. Here is a representative statement of the Theory Theory from Gopnik & Andrew Meltzoff:

> 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. [8]

The Theory Theory has also been defended in the literature on autism. A large body of evidence indicates that autistic individuals have severe deficiencies in their understanding of other minds, and this has led researchers to propose analogous deficiencies in autistic individuals' understanding of their own mental states.[9] Uta Frith and Francesca Happé express

the view as follows:

> …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[10].

Theory Theorists haven't been sufficiently clear about exactly how the Theory of Mind fits into the rest of the process of self-awareness.[11] But what is clear is that they regard Theory of Mind as a necessary component of self-awareness. All access to one's own mental states is "theoretical" in the sense that it depends on the Theory of Mind. So one cannot detect one's own mental states without exploiting the Theory of Mind that is used for detecting others' mental states.

In response to this growing consensus, Stich and I argued, rather, that the mind contains a "Monitoring Mechanism", a special purpose mechanism (or set of mechanisms) for detecting one's own mental states, and this mechanism is quite independent from the mechanisms that are used to detect the mental states of others. On the theory we develop, the Monitoring Mechanism (MM) takes as input one's own mental state (e.g., a belief, desire, or intention) and produces as output the belief that one has that mental state. So, for instance, if one believes that *p* and the Monitoring Mechanism is activated (in the right way), it takes the representation *p* in the Belief Box and produces the belief *I believe that p.* (See figure 1). This mechanism is

computationally extremely simple. For instance, to produce representations of one's own desires, the MM simply copies a representation from the Desire Box, embeds the copy in a representation schema of the form: *I desire that* ___., and then inserts this new representation into the Belief Box. Our proposal, then, was that the Monitoring Mechanism is an independent introspection mechanism for detecting one's own beliefs, desires, intentions and imaginings.[12]

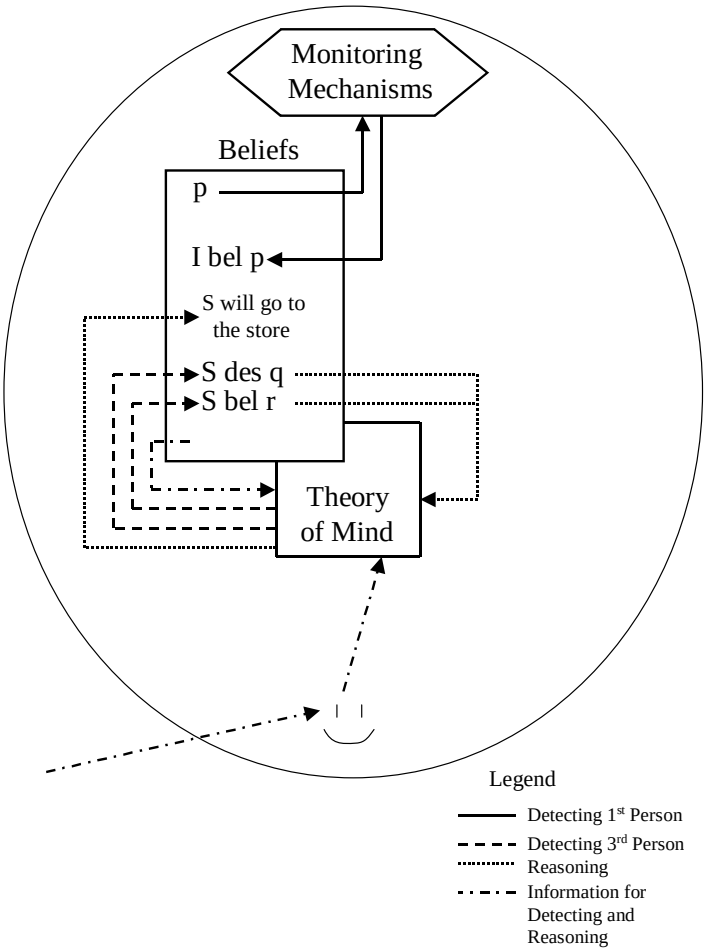

Figure 1: Monitoring mechanism theory of self-awareness

Although we argue that the Monitoring Mechanism theory is a much more plausible account of self-awareness than the Theory Theory, we do not deny that one can use the Theory of Mind on oneself. Indeed, we maintain that Theory of Mind is probably required for

reasoning about one's own mental states, i.e., using information about one's own mental states to predict and explain one's own mental states and behavior. However, it is quite a different matter for *detecting* one's own mental states. The detection of one's own mental states does *not* depend on the capacity for detecting or reasoning about other people's mental states. Thus, on our account there are special introspection mechanisms for detecting one's own mental states but not for reasoning about one's own mental states.[13]

Evidence

Theory Theorists have put forth a number of empirical arguments for the Theory Theory of self awareness. Stich and I provide detailed responses to these arguments and offer a few of our own arguments against the Theory Theory.[14] I won't rehearse all of those arguments here, but I do want to review briefly one of the arguments against the Theory Theory and then offer a couple of new arguments.

Development asynchronies

The most explicit and carefully charted argument for the Theory Theory comes from Gopnik & Meltzoff.[15] According to the Theory Theory, self-attributions will be subject to the same deficiencies as other-attributions. Hence, young children's mistakes on attributing mental states to others should find parallels in self-attribution. Gopnik & Meltzoff maintain that in fact children's understanding of their own mental states *does* develop in close parallel with their understanding of others' mental states. However, a closer look at the data suggests that the developmental evidence actually poses a problem for the Theory Theory. For on a wide range of tasks, children do not exhibit the parallel performance predicted by the Theory Theory.

Children are capable of attributing knowledge and ignorance to themselves before they are capable of attributing those states to others; they are capable of attributing certain perceptual states to themselves before they are capable of attributing such states to others; there is even some evidence that children are capable of attributing false beliefs to themselves before they are capable of attributing such states to others.[16] Hence, although Gopnik & Meltzoff claim that the developmental evidence supports the Theory Theory of self awareness, the evidence actually seems to undermine the Theory Theory in a fairly serious way. The Monitoring Mechanism account easily accommodates the data, however. The Monitoring Mechanism is proposed as an innate and early emerging mechanism, and the account does not predict that the capacity to detect one's own mental states will develop in parallel with the capacity to detect mental states of others.

Egocentric attributions

A further argument against the Theory Theory emerges from the developmental data when one considers the *kinds* of mistakes that toddlers make about other minds. When asked what another person, the "target", thinks or wants, toddlers do not respond at chance. Rather, for an important class of cases, they tend to attribute their own mental states "egocentrically". Indeed, this is how *Theory Theorists* characterize the mistakes. For instance, in one task, children are told to hide an object from the experimenter, and young 2 year olds failed this task. Gopnik & Meltzoff describe the young children's mistakes as follows: "24-month-olds consistently hid the object egocentrically, either placing it on the experimenter's side of the screen or holding it to themselves so that neither they nor the experimenter could see it."[17] Similarly, Repacholi & Gopnik found that 14-month old children shared the kind of food they

themselves liked rather than the food that the target exhibited a preference for. The experimenter made a facial expression of disgust or happiness after tasting either Goldfish crackers or broccoli. Although the 14-month olds consistently shared the crackers (their own preference), the 18 month olds were sensitive to the facial expressions of preference. Repacholi & Gopnik suggest the possibility that although the 14-month olds "were beginning to acquire a psychological conception of desire, it was, nonetheless, egocentric. Thus, although they understood that people request things because of some underlying desire, they mistakenly believe that everyone's desires are the same." [18] Elsewhere, Meltzoff, Gopnik & Repacholi suggest that the performance of 18 month olds on this task indicates that they have a more developed Theory of Mind: "This is a developmental achievement inasmuch as 14-month-olds did not do this. Instead, they always gave the experimenter crackers, their own preference, regardless of the experimenter's expressed desires. This work suggests that even very young children, 18-month-olds, may have a nonegocentric understanding of the differences between their own mental states and those of others in some cases." [19] These experimental findings of egocentric desire attributions are corroborated by ecological reports.[20] For instance, when young children help others in distress they tend to offer their own comfort objects (e.g., their teddy bear or blanket) to the distressed person.[21]

For the Monitoring Mechanism theory, early egocentric errors in mindreading pose no problem. Even before the young child has an adequate theory of, say, desire, she can use the Monitoring Mechanism to determine her own desires and preferences, and she can subsequently attribute her preferences to a target. By contrast, it is hard to see how a Theory Theorist can accommodate egocentric attributions. For if the child has a deficient Theory of Mind, such that she is incapable of detecting the desires of others, then the Theory Theory predicts that she

should also be incapable of detecting her own desires. Egocentric mistakes indicate an asynchrony – the young child is apparently aware of her own mental states and attributes them to others before she is capable of detecting the other person's distinctive mental states. Without further explanation, it is difficult to see how a Theory Theorist can consistently maintain both that toddlers have an early egocentric Theory of Mind and that one's access to one's own mental states depends on the same theory that is used to detect the mental states of others.[22]

Dissociations in psychopathologies

In addition to appeals to developmental evidence, a number of philosophers and psychologists have recently argued that the Theory Theory of self awareness is supported by psychopathological evidence on autism and schizophrenia.[23] The data on autism are of particular significance for Theory Theorists, since autistic children are widely regarded as having deficient mindreading capacities. For instance, autistic children continue to fail the false belief task well after their mental age peers pass the task.. And studies of spontaneous speech indicates that autistic children basically never talk about cognitive mental states like beliefs or thoughts. [24] Given the Theory of Mind deficit in autism, the Theory Theory predicts that autistic children should be similarly impaired at detecting their own mental states. Carruthers and Frith & Happé argue that case studies indicate that autistic individuals *do* lack access to their own mental states.[25] However, a close inspection of the evidence tends to undermine rather than support the Theory Theory. [26] For instance, in a diverse range of case studies, autistic individuals report their own mental states much better than Theory Theorists predicted. The Monitoring Mechanism theory has a ready explanation for this: the Monitoring Mechanism might be intact in autism despite the deficit to Theory of Mind. [27]

Although the evidence adduced by Carruthers and Frith & Happé seems to provide better evidence *against* the Theory Theory rather than for it, the evidence is in any case rather fragmentary, relying largely on case studies. There is new experimental evidence, however, that further confirms the claim that the Monitoring Mechanism is intact in autism despite the problems with Theory of Mind. In a recent set of studies, Farrant and colleagues found that autistic children did remarkably well on "metamemory" tests.[28] In metamemory tasks, subjects are asked to memorize a set of items and subsequently to report on the strategies they used to remember the items. In light of arguments from Theory Theorists, the experimenters expected autistic children to perform much worse than non-autistic children on metamemory tasks: "On the basis of evidence that children with autism are delayed in passing false belief tasks and on the basis of arguments that mentalizing and metacognition involve related processes, we predicted that children with autism would show impaired performance relative to controls on false belief tasks and on metamemory tasks and that children's performances on the two types of task would be related."[29] However, contrary to the researchers' predictions, there was no significant difference between the performance of autistic children and non-autistic children on a range of metamemory tasks. In one task, the subject was asked to remember a set of numbers that were given. The children were subsequently asked "What did you do to help you to remember all the numbers that I said?". Like the other children in the study, most of the autistic children gave some explanation that fit into the categories of 'thinking', 'listening' or 'strategies'. For instance, one autistic child said "I did 68, then the rest, instead of being six, eight, you put 68." Indeed, Farrant et al. claim that it is clear from the data that "there was no relation between passing/failing false belief tasks and the categories of response given to the metamemory question."[30] Although the results flouted the experimenters' Theory-Theory-

based prediction, they fit perfectly with the Monitoring Mechanism theory. For the Monitoring Mechanism can be intact even when Theory of Mind is damaged.

The Introspection Module

Thus, the Monitoring Mechanism theory fits the available evidence much better than the Theory Theory. Of course, historically, the most venerable account of introspection is not the Theory Theory (which is likely an invention of the 20th century), but rather that introspection is a species of perception, inner perception. On traditional inner perception models, one detects one's own mental states via experience or phenomenological features.[31] If the perception-model of introspection is developed in this way, then the MM theory diverges in an important way. For the Monitoring Mechanism does not rely on *phenomenological* features for identifying one's beliefs and desires. [32] In that sense, on the Monitoring Mechanism account, introspection is quite different from perception. However, there is an important way in which the Monitoring Mechanism might be akin to perception: it's plausible that both systems are modularized.

Perceptual systems are the paradigm examples of modules, as modularity is developed in Fodor's classic treatment in *The Modularity of Mind.* The central feature of modularity for Fodor is informational encapsulation. [33] A cognitive mechanism is encapsulated if it has little or no access to information outside of its own proprietary database. Perceptual systems tend to be encapsulated – there are restrictions on the kinds of information that are processed by perceptual systems. The classic illustration of perceptual encapsulation is the fact that the Müller-Lyer illusion persists even after one knows about the illusion. Apparently, the perceptual system is insensitive to the knowledge that the lines are the same length. Fodor

maintains that there are a number of other features that tend to co-occur with encapsulation. And Fodor maintains that perceptual systems also have these correlated features. Among other things, perceptual systems tend to be dedicated to particular tasks, they have characteristic ontogenies, and they exhibit characteristic patterns of breakdown, and perceptual processing tends to be very fast.

The Monitoring Mechanism, like the perceptual systems, has many of the features of modules. Like perceptual systems, the Monitoring Mechanism is dedicated to a particular task, it has a characteristic and early ontogeny, it seems to exhibit a characteristic pattern of breakdown[34], and it seems to be selectively spared in autism. It also exhibits fast, but restricted processing. The last point is of some significance. The processing capacity of the Monitoring Mechanism is extremely limited – it simply plugs a representation into the self-attribution schema. Thus, while it may not be strictly speaking encapsulated, the Monitoring Mechanism resembles encapsulated mechanisms in that it does not engage in any remotely intelligent general-purpose reasoning. We might, then, think of the Monitoring Mechanism as the *Introspection Module*.[35]

Possible Functions of Introspection

The evidence suggests, then, that the capacity for detecting one's own mental states is subserved by a Monitoring Mechanism that is quite independent from Theory of Mind. One might wonder, though, *why* we would have such a mechanism.[36] The problem is especially acute if one assumes that the Monitoring Mechanism evolved before Theory of Mind. As noted above, the developmental evidence indicates that the Monitoring Mechanism emerges earlier than Theory of Mind. So one might suppose that, if ontogeny recapitulates phylogeny, then the

Monitoring Mechanism should be phylogenetically older than Theory of Mind. Of course, the claim that ontogeny recapitulates phylogeny is hardly a strict law.[37] So it is *possible* that the Monitoring Mechanism emerged after or contemporaneously with Theory of Mind. Nonetheless, the claim that ontogeny recapitulates phylogeny has been a good heuristic, and it's probably better to be on the side of the heuristic rather than against it.

The clearest way to develop and confirm evolutionary accounts of cognitive mechanisms is by appeal to evidence comparing distantly related species that share a cognitive mechanism or related species in which the cognitive mechanisms diverge.[38] Unfortunately, in the case of introspection, we have no comparative evidence that speaks to this issue directly, so it's difficult to identify the evolutionary function of the Monitoring Mechanism with any confidence. However, to forestall the criticisms that such a mechanism would have been good for nothing, let me make clear a couple of different ways in which the Monitoring Mechanism could have been adaptive.[39]

One of the arguments I presented above against the Theory Theory is the fact that young children tend to attribute their own mental states *egocentrically*. Although egocentric attribution typically shows up most clearly in the mistakes that children (and adults[40]) make, in fact, the practice of egocentric attribution likely forms a large and productive part of our mindreading abilities. In trying to figure out another person's mental states, one quite successful strategy for a large set of states is to attribute one's own mental states to the target. For we typically share a broad background of similar beliefs with those we attribute beliefs to. This is also true for a wide range of preferences – my conspecifics and I typically have similar tastes and basic desires. Indeed, even in the Goldfish crackers & broccoli experiment, it's likely that the young infants who "mistakenly" attribute their own preferences to the target are adopting a fairly

effective strategy. Thus, a mechanism that provides access to one's own mental states might provide a basis for attributing mental states to others. And there are lots of reasons to think that it's adaptive to be good at attributing mental states to others.[41]

In a quite different way, monitoring one's own mental states might have been useful to enable more efficient planning. Perhaps the most influential account of planning in recent philosophy and artificial intelligence is Michael Bratman's "planning theory of intention."[42] Although it is not entirely explicit in Bratman's planning theory, the capacity to detect one's own mental states can play a crucial role in enabling efficient planning, as I hope to explain.

Since there are always indefinitely many possible courses of action and the world is constantly changing, ideal practical reasoning is quite impossible. It would require a constant assessment of the best thing to do, and humans have neither the leisure nor the capacity for such endless calculation. Bratman argues that one way out of this problem for a resource-limited creature is by *committing* to an intention in such a way that one no longer deliberates *all things considered* about what to do. Rather, one's commitment to an intention constrains and structures subsequent planning and decision making. One crucial part of the theory is that once you commit to a plan, there is a wide range of incompatible options that you don't even consider – they are "filtered" out.[43] For instance, if I commit to going to England on July 15th, then I don't even consider the option of having a dinner party on July 16th. Of course, this deliberative neglect of a range of options carries a certain cost – for some of the options that aren't considered might have been adopted had the options been considered in a thorough process of deliberation. However, in many other cases, the options that are neglected would not have been adopted in any case, so in those (presumably, more typical) cases, one saves the time and energy of deliberation without incurring any costs. The other crucial feature of committing

to an intention is that this structures one's subsequent deliberations, e.g., about the best way to execute the plan. So, once I'm committed to going to England, I generate the subplan that I need to get my passport renewed. In the long run, this filtering and structuring plausibly makes one's planning much more efficient than a constant calculation of utility maximization.

In her work in Artificial Intelligence, Martha Pollack has argued that another way that resource-bounded agents can make their reasoning more efficient is by "overloading" their intentions.[44] The idea is that if an agent has a goal, she can try to determine whether that goal can be satisfied in the course of executing some plan that has already been adopted. So, for instance, suppose I realize that I need to buy a present for a party that will occur tomorrow. I can then consider whether I already have a plan that will take me near an appropriate vendor. If I had already planned to go downtown to the post office, I can "overload" this prior plan to include a trip to the store. Not only will this make the errand-running more efficient, Pollack suggests that by overloading one's intentions in this way, one's reasoning is also more efficient.

How does all this connect to the utility of the Monitoring Mechanism? What I want to suggest is that overloading or committing to one's intentions is enabled by having a mechanism that delivers beliefs about one's intentions. Since Bratman and Pollack develop their theories in contexts in which it is simply assumed that the reasoning agents (be they human or artificial) can represent their own intentions, this part of the theory is never made explicit. However, unless one knows about one's own intentions, it's difficult to see how one can overload one's intentions. Similarly if you don't know which intentions you're "committed" to, it's hard to see how those intentions can subsequently inform one's planning. The obvious way to implement the kind of reasoning that Bratman & Pollack promote would involve having *beliefs* about your own intentions. If the agent has a *belief* about what she intends to do, this belief can structure

her further decision making. Of course, the Monitoring Mechanism would serve the function of delivering such beliefs. Hence, it seems that this mechanism might play a crucial role in facilitating efficient planning. Again, I'm not arguing that the Monitoring Mechanism actually evolved to serve this function. To make such a claim plausible one would need a body of comparative evidence. Rather, the point is to show that there could be fairly direct advantages to having a Monitoring Mechanism, so the mechanism can't be faulted on general evolutionary grounds.

To summarize this section, I've offered a number of arguments for why the Monitoring Mechanism account of introspection is more plausible than the rival Theory Theory account. I've also proposed some reasons to think that such a mechanism could have been adaptive. And I've suggested that this Mechanism is modular. There certainly is not sufficient evidence for this claim to parade it as an obvious truth. But given the available evidence, it is a plausible conjecture that there is a modular Monitoring Mechanism – a dedicated mechanism for detecting one's own mental states that is independent of the capacity to detect mental states of others.

III. THE SELF-CONCEPT IN COGNITIVE SCIENCE

If the foregoing account of introspection is right, it might have significant implications for psychological work on the concept of self.[45] For the Monitoring Mechanism account suggests that there is a basic concept of self that has been largely neglected in the psychological literature. There is a vast literature on the self in psychology elaborating, often in great detail, how people think of themselves. Much of this work focuses on issues that relate to self-esteem

and self-worth.[46] But here I want to focus on the ontogeny of "self-conception", i.e., when children develop a concept of self. Perhaps the best known method for addressing this question is the mirror self-recognition task.[47] More recently, in light of the research on mindreading, researchers have suggested rather that the child's concept of self depends on the Theory of Mind.[48] If the Monitoring Mechanism theory is right, then there is a notion of self that isn't captured by either of these approaches.

The Body's "I"

There is a long tradition in developmental psychology and primatology of using "mirror self-recognition" tests to descry awareness of self in infants and nonlinguistic animals. The first study on mirror self-recognition was reported by Gordon Gallup.[49] Gallup found that chimpanzees respond in self-exploratory ways to their images in mirrors; for example, they inspect parts of their body that are difficult to see without a mirror, and they will reach up to investigate marks that were surreptitiously placed on the forehead before the mirror was made available. Subsequent research on human children showed that children begin to exhibit this kind of behavior by 18-24 months. [50] A wide range of further comparative research has found, surprisingly, that most species do *not* exhibit this kind of self-exploratory behavior. For instance, it has not been convincingly demonstrated in any species of monkey.[51]

Although mirror self-recognition is sometimes treated as evidence that chimpanzees have a psychological understanding of themselves[52], there are fairly obvious alternative explanations. For instance, Daniel Povinelli maintains that passing the mirror tasks doesn't require a concept of self as psychological subject: "Gallup believes that chimpanzees possess a psychological understanding of themselves. In contrast, I believe these apes possess an explicit

mental representation of the position and movement of their own bodies – what could be called a kinesthetic self-concept."[53] On Povinelli's account it is the "contingency between the self's actions and the actions in the mirror" that "triggers the formation of an equivalence relation between the organism's internal self-representation and the external stimuli (the mirror image)".[54] In fact, as Povinelli points out, this is similar to the proposal Gallup made in his initial paper reporting the findings. "Gallup's (1970) initial speculation was that 'self-directed and mark-directed behaviors would seem to require the ability to project, as it were, proprioceptive information and kinesthetic feedback onto the reflected visual image so as to coordinate the appropriate visually guided movements via the mirror' (p. 87)".[55] So, while chimpanzee & human toddler's self-exploratory behavior in front of mirrors is plausibly evidence that chimpanzees and toddlers have a concept of the self-as-body, the *Body's "I"*, it does not show that they have the concept of self as a mind. The capacity for kinesthetic feedback may suffice for generating mirror self-recognition.[56]

What does the absence of self-exploratory behaviors in front of mirrors show? It does not show that the creature lacks any concept of self. Indeed, it may well be the case that creatures can "fail" the mirror task while having a psychological understanding of the self. If, for instance, Povinelli's theory of mirror self-recognition is right, then a system of kinesthetic monitoring plays a crucial role in mirror self-recognition. As a result, creatures that lack this kinesthetic monitoring might well fail the mirror self-recognition task. But it's quite possible that a creature can have an understanding of the self-as-mind without kinesthetic monitoring. A subject that lacked any kinesthetic monitoring might know that people (including himself) have beliefs and desires without being able to detect the contingencies between his bodily movement and that of his mirror image. Indeed, the possibility of a dissociation between mirror self-

recognition and self-awareness has recently been given some empirical support. Breen and colleagues report cases in which subjects have deficits in mirror self-recognition (i.e., they do not recognize themselves in mirrors) but apparently no deficit in self-awareness.[57]

The upshot is that while mirror self-recognition provides evidence of some concept of self, the mirror tasks don't tell us much about the subject's understanding of herself as a subject of psychological states. It's possible to exhibit mirror self-recognition while having no understanding of the self-as-mind, and it's possible that creatures can have an understanding of the self-as-mind even if they don't exhibit mirror self-recognition.

*The **Theory of Mind**'s 'I'*

Although mirror tasks seem to provide no evidence that a creature understands itself as having psychological states, the work in Theory of Mind provides a wealth of evidence on this in humans. There is little question now that pre-school children understand that they (and others) have psychological states. Thus, the child's Theory of Mind obviously provides the basis for a psychological understanding of the self.

Unfortunately, the task of characterizing the young child's concept of self has not yet received sustained attention by researchers drawing on the Theory of Mind tradition. However, the view that the concept of self depends on Theory of Mind has been suggested by various authors in different ways.[58] It's also implicit in the Theory Theory of self awareness, since on that theory, the child's understanding of her own psychological states depends on the Theory of Mind. Perhaps the most explicit statement that the concept of self depends on Theory of Mind comes from Henry Wellman:

> our understanding of ourselves partakes of and is limited by our framework belief-desire psychology…. Thoughts, desires, basic emotions, actions, perceptions and so on, as specified by belief-desire psychology, are the basic categories that frame specific person concepts. Everyday theory of mind provides the infrastructure for self-conception.[59]

Wellman doesn't directly consider mirror self-recognition, but it's reasonable to assume that he would, in line with the discussion above, maintain that mirror self-recognition does not show anything about the child's understanding of the self as a psychological subject. Once we put the concept of self-as-body to the side, Wellman seems to be suggesting that the child's concept of self depends crucially on the Theory of Mind. We might call this concept the *Theory of Mind's 'I'*.

As noted above, there is disappointingly little written in this tradition on the child's concept of self. Earlier developmental work on the child's concept of self produced the bizarre finding that children seemed to identify the self primarily with physical features in free recall tests.[60] For example, when young children are asked "What will (not) change about yourself when you grow up?", 7 year olds tended to refer to physical characteristics (e.g., hair color) and very seldom referred to psychological characteristics.[61] Indeed, one prominent view was that young children had only a "physicalistic" conception of the self. [62] This body of findings seemed increasingly peculiar in light of the emerging body of evidence on the young child's extraordinary capacities in Theory of Mind. The research on Theory of Mind clearly demonstrates that young children attain considerable sophistication about the mind well before the age of 7. Hence, it's puzzling that they should identify themselves solely with their physical features.

In a set of recent experiments motivated partly by these considerations, Daniel Hart and

colleagues found that although children tended to appeal to their physical features in standard free-recall tasks about the self, the same children regard their psychological characteristics as most important to their self.[63] Hart and colleagues used philosophically-inspired thought experiments on personal identity to explore this. In one condition, the child is shown a model of a "person machine" and is told the following:

> This is a person machine. What the machine does is make persons. The person behind this door gets an exact copy of your body and looks exactly like you. But this person… does not have your thoughts and feelings…. The person behind this door… has an exact copy of your thoughts and feelings… But this person does not have your body or look like you.[64]

The subject was then asked, "Which of these two persons is closest to being you? The one with your body and appearance or the one with your thoughts and feelings?" The researchers compared children's responses to the person machine task with their responses to standard free-recall tasks, and found that in the free recall tasks, the children mentioned physical characteristics most often, but in the person machine task, they regarded their psychological qualities as most important for the self. Hart and colleagues describe the results as follows:

> When asked to judge which set of characteristics was most important for establishing similarity between the self and a hypothetical person, the 7-year-olds in this study most frequently claimed that it is their psychological features; indeed, over half of the children claimed that the psychological characteristics are superior for preserving personal identity for all the hypothetical transformations posed in this study.[65]

Philosophers will no doubt notice that this isn't really a question about personal *identity*, since it is explicitly about similarity between simultaneously existing persons. However, many children

would be upset by imagining that they are dismantled shortly after stepping into the person machine, and ethics review boards would be unlikely to approve such a study should a sadistic experimenter propose it. At any rate, as Hart and colleagues intimate, the person machine task is likely more revealing than free-recall tasks for uncovering the child's theory of the core features of the self.

Hart and colleagues developed their task explicitly in the context of Theory of Mind research. And the Theory of Mind is plausibly implicated in the person machine task since the task requires the subject to judge the importance of psychological properties for similarity across individuals. The findings on the person machine task begin to tell us a bit about the Theory of Mind's "I", then. They suggest that on the concept of self delivered by the Theory of Mind, the most important features of the self are one's psychological properties.[66] Presumably children don't have this understanding of self until they have a Theory of Mind. Hence, at least for this concept of self, Wellman is right: "Everyday theory of mind provides the infrastructure for self-conception."

The Mind's "I"

So, the Theory of Mind apparently delivers a concept of self as a mind, and this concept needs to be distinguished from the concept of self-as-body. However, the Monitoring Mechanism account suggests that there is another, more basic concept of self that is independent of Theory of Mind, what one might call the *Mind*'s "I".

The Monitoring Mechanism sketched in section II produces beliefs about one's own mental states. That is, it produces representations of the form *I believe p, I desire q, I imagine r,* etc. Thus, as the source of such self-attribution, this mechanism delivers representations in

which the concept, *I,* is the subject of mental states. Since the Monitoring Mechanism is presumed to be innate, one of the implications of the account is that this concept of the self as subject is also innate.

The Mind's "I" is distinct from both of the self-concepts discussed above. It's distinct from the Body's "I" since the Mind's "I" is explicitly the subject of psychological states. There is no *a priori* reason to think that this concept of self will covary with the concept of self that is revealed through mirror self-recognition. It's certainly possible that the concept of the Mind's "I" is present in creatures that lack the concept of the Body's "I". For instance, if Povinelli is right, mirror self-recognition depends on a keen system of kinesthetic monitoring in humans and chimpanzees. But of course, the evolutionary pressures that led to a keen system of kinesthetic monitoring may well be quite different from those that led to mental state monitoring. And, as noted earlier, a creature might well be able to recognize its own mental states without recognizing the contingencies between its movements and the movements in the mirror.[67]

Thus, the Monitoring Mechanism delivers a concept of the self as a mind, rather than a body. However, this is a concept of self-as-mind that does not depend on Theory of Mind. Rather, it is present in humans before Theory of Mind has matured, and it may be present in creatures that don't *have* Theory of Mind. In contrast to Wellman's claim, then, there is a basic concept of the self as a psychological subject for which the Theory of Mind does not provide the infrastructure. And while the concept of self that depends on Theory of Mind may well exhibit cross-cultural variation[68], the Mind's "I" is unlikely to be cross-culturally variable for it is a direct output of an innate and early emerging module.

It's worth emphasizing that on the Monitoring Mechanism account, the fact that this

basic concept of "I" is delivered by introspection in no way suggests that there is a phenomenologically salient sense of the self as psychological subject. That is, this basic concept of the self as a psychological subject comes from a specialized computational module, not from any phenomenological features that the self might have.[69]

Thus far, I've focused largely on what this concept of self is not; it is harder to provide a positive characterization of the concept since psychologists have not studied it systematically. By hypothesis, one of the functional properties of this concept of self is to underwrite self attribution, and the concept is exploited by young children's early attributions of mental states to themselves; this concept is presumably also connected to action systems. But what does this self-concept specify as the core features of the self? It's likely that this concept of self contains virtually no information about the essential features of the self. As we've seen, the concept of self that depends on Theory of Mind seems to provide a rich characterization of the essential features of the self on which a person's psychological properties are crucial to the self. By contrast, the computational characteristics of the Monitoring Mechanism suggest that, while this simple mechanism does deliver a concept of self, this concept does not include any specification of the core features of the self.

Despite its exiguous content, this concept of self *can* support judgments of personal identity. I currently desire that it not rain, since I've discovered that my roof is leaking. And the Monitoring Mechanism can detect this desire. However, before I discovered the leak, I desired extensive rain to relieve the dustbowl that is my backyard. So, several weeks ago, the Monitoring Mechanism produced the belief that I want it to rain, and that belief was converted into the memory that I wanted it to rain. More recently, my desires have changed, and the Monitoring Mechanism produces the belief that I currently desire that it not rain. So, the

Monitoring Mechanism is the basis for the representations *I wanted it to rain* and *I want it not to rain*. And it's plausible that I can use such beliefs as the basis for the judgment that although I wanted rain previously I do not want it now. Of course, I can have several mental states at once, and the Monitoring Mechanism might deliver, e.g., the belief that I currently desire to go to the store and to listen to music. The upshot of this is that the Monitoring Mechanism provides the basis for making judgments of personal identity both synchronically and diachronically. I am the same person who wanted it to rain several weeks ago and currently wants it not to rain. I am also the same person who wants to go to the store and to listen to music.[70]

Although it hasn't been systematically studied, it's likely that these kinds of judgments of identity are implicit in self-attributions of young children. Consider, for instance, the following exchanges from the CHILDES database:

Abe (3;3): I didn't get you a surprise.

Adult: You didn't. I'm sad.

Abe: No, don't be sad. I thought I would 'cept I didn't see one for you.

Adult: Do you remember [what you got as a present]?

Abe (3;3): A net. [a basketball hoop]

Adult: Uh huh.

Abe: 'Cept I didn't want it. I wanted a bat and baseball.

Adult: I thought it was a bus.

Adam (3;3): It's a bus. I thought a taxi.[71]

In all of these cases, the child seems to make an implicit judgment of identity across time. For instance, Abe reports that he previously thought that he would get his father a surprise, but that he didn't see one. Similarly, in the experimental work on theory of mind and self attributions, young children seem to make these kinds of implicit identity judgments. For instance, in an experiment by Gopnik and Virginia Slaughter, the subject is shown a drawing of a turtle and then switches seats with the experimenter and asked "before we traded seats, how did you see the turtle then, lying on his back or standing on his feet?".[72] The young child reports that before changing seats, she saw the turtle differently than she does now. And in recent work by Tim German and Alan Leslie, the young child attributes a past false belief to himself to explain his past action, saying that he looked for the bait in the wrong box "because that's where I thought it was." In these cases as well, the child seems to identify herself with a subject of a past psychological state that differs from her current psychological state.

Some might maintain that this early concept of self is so content-poor that it cannot be regarded as a genuine concept of self. Rather, perhaps we should say that the Monitoring Mechanism delivers a "proto-concept" of self, and that it is only after the child can also exploit the Theory of Mind that she can be said to have a *genuine* concept of self.[73] I don't know how to decide these questions about the genuineness of concepts, but for present purposes I don't think it matters much, and I mean to use the term "concept" broadly enough to include such "proto-concepts". What is crucial is that the Monitoring Mechanism delivers an "I" which, whether a *genuine* concept of self or not, suffices for self-attribution and for judgments of personal identity across time.

The Monitoring Mechanism account thus suggests that there is a basic concept of self as a psychological subject that has been largely neglected in the psychological literature. The mirror self-recognition tasks don't show understanding of the self as a psychological subject. A large body of evidence in the Theory of Mind tradition does show that children have some psychological understanding of the self, and recent research indicates that psychological properties are regarded as central to the self by children. However, the Monitoring Mechanism account suggests an earlier innate notion of the self on which the self is subject of psychological states. This concept of self is not shown by the mirror self-recognition tasks, nor is it dependent on the Theory of Mind.

IV. THE SELF-CONCEPT IN PHILOSOPHY

One of the enduring issues in philosophy of mind and metaphysics has been the attempt to characterize the conditions under which a person at one time is identical to a person at another time. Of course, one of the reasons this project is of interest is that by characterizing the conditions for personal identity, one thereby characterizes the essential features of the self. The problem of personal identity has seemed to many to be one of the deepest and most difficult problems in philosophy of mind and metaphysics. The difficulty is that it has seemed impossible to develop an account of the self and personal identity that coheres with all of our intuitions. Indeed, Colin McGinn has recently argued that the reason the problem is so perplexing is that we lack the cognitive resources to solve it.[74] I want to suggest a rather different kind of explanation for why we are unable to reconcile our intuitions about the self. The deeply puzzling problems arise, I'll suggest, because we have two quite different cognitive

mechanisms that generate judgments about the self.

The problem of personal identity is often characterized as a conflict between intuitions issuing from the 3rd person perspective and intuitions issuing from a 1st person perspective.[75] The point is nicely put by Thomas Nagel:

> The problem [of personal identity] seems unreal when persons are viewed as beings in the world, whether physical or mental. They persist and change through time, and those are the terms in which they must be described… the persistent dissatisfaction with candidate analyses of this form derives from a submerged internal aspect of the problem which is left untouched by all external treatments. From the point of view of the person himself, the question of his identity or nonidentity with someone undergoing some experience in the future appears to have a content that cannot be exhausted by any account in terms of memory, similarity of character, or physical continuity. Such analyses are never sufficient, and from this point of view they may appear not even to supply necessary conditions for identity.[76]

I think that the psychological work on the self concept suggests an explanation for the conflict that Nagel presents. The idea is that, in general, the 3rd person perspective on the self recruits the Theory of Mind and the 1st person perspective recruits the Monitoring Mechanism; as a result, different concepts of the self are implicated from the different perspectives. That is, the 3rd person perspective and 1st person perspective thought experiments recruit different cognitive mechanisms that exploit different concepts of the self.[77] In the remainder of this section, I'll try to clarify this suggestion.

In the philosophical literature on the self, one of the best known thought experiments from the 3rd person perspective comes from Bernard Williams.[78] Drawing from a classic

Lockean example, Williams asks us to imagine that a mad scientist will perform an operation on two persons, A and B. Each patient will have his psychological properties (thoughts, feelings, personality traits) "removed" from his brain and transferred into the brain of the other person. As a result, the A-body person will have the psychological properties that B had before the switch, and the B-body person will have the psychological properties that A had before the switch. Before the operation, the procedure is explained to A and B, and they are told that after the transfer, one of the resulting persons will be tortured and the other will be given a large sum of money; A and B are then allowed to request which body get the torture and which the money (of course, it may be impossible for the mad scientist to grant both of their requests). After the operation, the subjects are asked whether they got what they requested.

Beginning with Locke's initial presentation of these sorts of cases, most philosophers have the intuition that this seems to be a case of body-swapping – in virtue of the transfer of their psychological properties, A and B have switched bodies.[79] Williams makes this seem particularly compelling by considering the range of possible responses A and B would make before and after the transfer. For the B-body person will remember A's request, and as a result, the B-body person will say he got what he asked for if and only if A's request was met (and the situation is analogous for the A-body person). Thus, Williams maintains that from the 3rd person perspective it seems that to transfer psychological properties is to transfer the self. Thought experiments like Williams' 3rd person case have led philosophers to develop accounts of the self according to which continuity of psychological properties is essential to personal identity.[80] Intuitions about these 3rd person cases presumably depend on Theory of Mind, since the thought experiments require fairly sophisticated reasoning about others' psychological properties. And the philosophical intuitions about these cases seem to fit with the available

psychological evidence on the concept of self that is delivered by the Theory of Mind. The evidence indicates that, at least in Western culture, psychological properties are regarded as essential features of the self. And the emerging evidence from developmental psychology indicates that children share this intuition. Indeed, the Lockean thought experiment is a direct ancestor of task used in the developmental work on the Theory of Mind-concept of self.

When we turn to the philosophical intuitions from the 1st person perspective, the situation is quite different. The problem is that from the 1st person perspective, continuity of psychological properties does not seem essential to personal identity; but neither does continuity of physical properties. Nagel puts the point as follows:

> The concept of the self seems suspiciously pure – too pure – when we look at it from inside…. When I consider my own individual life from inside, it seems that my existence in the future or the past – the existence of the same 'I' as this one – depends on nothing but itself....My nature then appears to be at least conceptually independent not only of bodily continuity but also of all other subjective mental conditions, such as memory and psychological similarity. The migration of the self from one body to another seems conceivable, even if it is not in fact possible. So does the persistence of the self over a total break in psychological continuity – as in the fantasy of reincarnation without memory. If all these things are really possible, I certainly can't be an organism: I must be a pure, featureless mental receptacle.[81]

I've suggested that the 3rd person perspective thought experiments exploit the Theory of Mind. When we take the 1st person perspective on the self, it's plausible that we are using the Monitoring Mechanism, since our distinctive access to our own minds is largely subserved by that mechanism. From the 3rd person perspective, psychological continuity seems to be crucial

to identity across time. But from the 1st person perspective, psychological continuity seems quite inessential to identity across time, as Nagel suggests – it's possible to imagine one's self existing after a radical disruption in psychological or physical properties. Just as the 1st person perspective thought experiments suggest that the self lacks essential connections to psychological or physical properties, we saw in section III that the concept of self delivered by the Monitoring Mechanism likely lacks any specifications of the nature of the self. The Monitoring Mechanism delivers a concept of self as subject of mental states, and young children exploit this mechanism to make judgments of identity across time. However, the concept of self delivered by the Monitoring Mechanism does not provide any characterization of the core features of the self. As a result, if our 1st person thought experiments recruit the Monitoring Mechanism and isolate this concept of self, presumably such thought experiments will not produce the intuition that continuity of psychological properties or physical properties is required for identity across time. Rather, if we try to characterize the self by relying on the Monitoring Mechanism and its attendant concept of self, we seem to arrive at something close to Nagel's characterization of the self from the 1st person perspective: "a pure, featureless mental receptacle."

If this is right, we have an explanation for why it has been so difficult to develop an account of the self that satisfies all of our intuitions about the self. The problem is that we have two sources of intuitions about the self. One source depends crucially on the Monitoring Mechanism; the other source depends on the Theory of Mind. These cognitive mechanisms exploit quite different self-concepts which generate incompatible intuitions about the nature of the self.[82] As a result, the philosophical task of developing a theory that can accommodate both the 3rd person and the 1st person intuitions about the self is likely to be hopeless.

This is not the place to defend a positive metaphysics of the self. But the above diagnosis of the problem suggests that we should be suspicious of intuitions about the self that issue from the 1st person perspective. For on the account I've suggested, the generation of these 1st person intuitions typically depends on a rather simple module, which exploits a quite minimal concept of self. And the function of the module is to inform the subject of her mental states, not to inform the subject of the nature of the self. A central characteristic of modules is that they are *very* good at a certain range of tasks, and very bad at pretty much everything else. Obviously, the Monitoring Mechanism is very good at detecting one's own mental states. In fact, the Monitoring Mechanism is much better at detecting one's own mental states than *any* other source, including Theory of Mind and contemporary scientific psychology. But, the Monitoring Mechanism is likely a very poor source for trying to build a philosophically viable account of the self.

V. CONCLUSION

Recent work on development and psychopathologies provides a tremendous resource for developing a cognitive account of introspection. This work suggests that there is a dedicated cognitive mechanism for detecting one's own mental states, a Monitoring Mechanism. I've argued that the psychological work also indicates that it's important to distinguish between different concepts of the self. In particular, the Monitoring Mechanism delivers a concept of self that needs to be distinguished from the richer concept of self that depends on the Theory of Mind. I've suggested that this distinction can begin to illuminate why the 1st person and 3rd person perspectives generate such deeply conflicting intuitions about the nature of the self. The problem is that the 1st person and 3rd person perspectives exploit different cognitive mechanisms

with different and incompatible concepts of the self. Obviously I have only offered the sketchiest defense of this proposal. But I hope to have shown that we should take seriously the possibility that a cognitive account of introspection might help us understand why the philosophical problem of the self has been so intractable.

NOTES

[1] I would like to thank Michael Gill, Alan Leslie, Martin Perlmutter, Philip Robbins, Eric Schwitzgebel, Lisa Shapiro, Steve Stich, and Matthew Stone for discussion and comments on an earlier draft of this paper.

[2] See, for example, Sydney Shoemaker, *The First-Person Perspective* (Cambridge: Cambridge University Press, 1996).

[3] For reviews of this literature, see Henry Wellman, *The Child's Theory of Mind* (Cambridge, MA: MIT Press, 1990); Josef Perner, *Understanding the Representational Mind* (Cambridge, MA: MIT Press, 1991); Alison Gopnik "How We Know Our Own Minds: The Illusion of First-Person Knowledge of Intentionality. *Behavioral and Brain Sciences*, 16 (1993): 1-14; Simon Baron-Cohen, *Mindblindness* (Cambridge, MA: MIT Press, 1995); Shaun Nichols and Stephen Stich, *Mindreading* (Oxford: Oxford University Press, forthcoming).

[4] Nichols and Stich, "How to Read Your Own Mind: A Cognitive Theory of Self-Consciousness," in *Aspects of Consciousness*, ed. Q. Smith and A. Jokic (Oxford: Oxford University Press, forthcoming) and "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, forthcoming).

[5] This view is advocated in several places: Simon Baron-Cohen, "Are autistic children 'behaviorists'?" *Journal of Autism and Developmental Disorders* 19 (1989): 579-600; Gopnik, "How We Know Our Own Minds", Alison Gopnik and Henry Wellman, "The Theory Theory," in *Mapping the Mind* ed. S. Gelman and L. Hirschfeld (Cambridge: Cambridge University Press, 1994); Alison Gopnik and Andrew Meltzoff, "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, 1994); Perner, *Understanding the Representational Mind*; Heinz Wimmer and Michael Hartl, "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 (1991):125-28; Peter Carruthers, "Autism as mind-blindness: An elaboration and partial defence" in *Theories of theories of mind*, eds. P. Carruthers and P. Smith. (Cambridge: Cambridge University Press, 1996); Christopher Frith, "Theory of Mind in Schizophrenia" in *The Neuropsychology of Schizophrenia*, ed. A. David and J. Cutting (Hillsdale, NJ: LEA, 1994); Uta Frith and Francesca Happé, "Theory of Mind and Self Consciousness: What Is It Like to Be Autistic? *Mind & Language*, 14 (1999): 1-22.

[6] Wilfrid Sellars, "Empiricism and the Philosophy of Mind," *Minnesota Studies in the Philosophy of Science*, vol. 1. (University of Minnesota Press, 1956). Reprinted in Sellars, *Science, Perception and Reality* (London: Routledge and Kegan Paul, 1963).

[7] J. Perner, S. Leekam, and H. Wimmer, "Three-year Olds' Difficulty with False Belief: The Case for a Conceptual Deficit," *British Journal of Experimental Child Psychology*, 39 (1987): 437-71; see also Wellman, *A Child's Theory of Mind*; Perner, *Understanding the Representational Mind*; Baron-Cohen, *Mindblindness;* Karen Bartsch and Henry Wellman, *Children Talk about the Mind* (Oxford: Oxford University Press, 1995); Gopnik & Wellman, "The Theory Theory"; Alan Leslie, "Pretending and Believing: Issues in the Theory of ToMM," *Cognition*, 50 (1994): 211-238.

[8] Gopnik and Meltzoff, "Minds, Bodies, and Persons", 168; see also Gopnik, "How We Know Our Own Minds".

[9] Uta Frith and Francesca Happé, "Theory of Mind and Self Consciousness", 7; see also Uta

Frith, *Autism: Explaining the Enigma* (Oxford: Blackwell, 1989); Baron-Cohen, "Are autistic children 'behaviorists'?"; Carruthers, "Autism as mind-blindness".

[10] Frith and Happé, "Theory of Mind and Self Consciousness", 7.

[11] Stich and I treat this issue at length in "How to Read Your Own Mind" and "Reading One's Own Mind".

[12] In addition to the Monitoring Mechanisms for detecting one's own propositional attitudes, Stich and I also maintain that there are analogous mechanisms, Percept-Monitoring Mechanisms, for detecting one's own perceptual states. For ease of exposition, I will focus mostly on the mechanisms for detecting propositional attitudes.

[13] Our claim is not that *all* instances of detecting one's own mental states are driven by MM. Sometimes, one might use Theory of Mind to determine one's beliefs and desires. More interestingly, there may be certain kinds of mental states that can only be detected through the use of theory (Nichols and Stich, *Mindreading*). For instance, you probably can't attribute *Schadenfreude*, resentment, or wishful thinking to yourself without having some theory about these kinds of mental states. We might think of these as 'thick' self-attributions. Of course, in some such cases, it might be that the MM still plays an important role in generating thin attributions that are essential to the thick attributions.

[14] Nichols & Stich, "How to Read Your Own Mind" and "Reading One's Own Mind".

[15] Gopnik & Meltzoff, "Minds, Bodies, and Persons".

[16] For data on self attributions of ignorance, see H. Wimmer, G. Hogrefe, and J. Perner, "Children's Understanding of Informational Access as a Source of Knowledge," *Child Development*, 59 (1988): 386-96; I also discuss these data in "Developmental Evidence and Introspection," *Behavioral and Brain Sciences*, 16 (1993): 64-65. For evidence on self

attributions of perceptual states, see Alison Gopnik and Virginia Slaughter, "Young Children's Understanding of Changes in Their Mental States," *Child Development*, 62 (1991): 98-110. Recent data on self attribution of belief is presented in Tim German & Alan Leslie, "Self-other differences in false belief: Recall versus reconstruction" (forthcoming). For further discussion of these findings see Nichols & Stich, "Reading One's Own Mind".

[17] Alison Gopnik and Andrew Meltzoff, *Words, Thoughts and Theories* (Cambridge, MA: MIT Press, 1997), 116.

[18] Betty Repacholi and Alison Gopnik "Early Understanding of Desires: Evidence from 14 and 18 Month Olds," *Developmental Psychology*, 33 (1997): 12-21, 18.

[19] Andrew Meltzoff, Alison Gopnik, and Betty Repacholi, "Toddlers' Understanding of Intentions, Desires, and Emotions: Explorations of the Dark Ages," in *Developing Theories of Intention: Social Understanding and Self-Control*, ed. P. Zelazo, J. Astington, and D. Olson (Mahwah, NJ: Lawrence Erlbaum Associates, 1999), 32.

[20] Although Theory Theorists concede that desire is attributed egocentrically, the situation with egocentric attribution of belief is more complicated. The failure on the false belief task certainly *looks* like egocentric attribution of one's own belief – the subject attributes to the other person the belief that she herself has. However, developmentalists have largely resisted this interpretation, claiming instead that the mistake is a form of *reality bias* (e.g., P. Mitchell, E. Robinson, J. Isaacs, and R. Nye, "Contamination in Reasoning about False Belief: An Instance of Realist Bias in Adults but not Children," *Cognition*, 59 (1996): 1-21) or is the product of an immature, "realist" theory of belief (e.g., Gopnik, "How We Know Our Own Minds"). Nonetheless, it's hard to find an argument *against* the claim that young children egocentrically attribute their own beliefs. And, in some recent studies, Trisha Folds-Bennett and I found that

children who fail the false belief task also show a statistically strong tendency to attribute their own probabilistic beliefs and guesses to others ("Egocentric Strategies in Young Children's Attributions of Mental States", in prep.).

[21] Martin Hoffman, "Development of Prosocial Motivation: Empathy and Guilt," in *Development of Prosocial Behavior*, ed. N. Eisenberg (New York: Academic Press, 1982), 281-313.

[22] Perhaps Theory Theorists can say that in order to detect your own mental states, you need a piece of Theory of Mind that is also required for detecting other mental states. But without saying more explicitly what is needed, this does not circumvent the problem of egocentric attribution. Indeed, one possibility is the very antithesis of the Theory Theory – that one's access to one's own states provides a crucial basis for attributing mental states to others.

[23] Carruthers, "Autism as mind-blindness"; Frith & Happé, "Theory of Mind and Self Consciousness"; Frith, "Theory of Mind in Schizophrenia".

[24] The initial findings on autism and false belief were reported in Simon Baron-Cohen, Alan Leslie, and Uta Frith, "Does the autistic child have a "theory of mind"?" *Cognition*, 21 (1985): 37-46. Spontaneous speech in autism is analyzed in Helen Tager-Flusberg, "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, 1993), 138-157.

[25] Carruthers, "Autism as mind-blindness"; Frith and Happé "Theory of Mind and Self Consciousness".

[26] Nichols and Stich "How to Read Your Own Mind".

[27] Ibid.

[28] A. Farrant, J. Boucher, M. Blades, "Metamemory in Children with Autism" *Child Development* 70 (1999): 107-131.

[29] Ibid. 108.

[30] Ibid. 118, 119.

[31] Locke provides a classic statement of this view, but for a recent version, see Alvin Goldman, "The Psychology of Folk Psychology," *Behavioral and Brain Sciences*, 16 (1993): 15-28.

[32] Nichols and Stich, "How to Read Your Own Mind".

[33] Jerry Fodor, *Modularity of Mind* (Cambridge, MA: MIT Press, 1983); *The Mind Doesn't Work That Way* (Cambridge, MA: MIT Press, 2000).

[34] In Nichols and Stich, "How to Read Your Own Mind", we argue that there is a bit of evidence suggesting that the Monitoring Mechanism is damaged in certain forms of schizophrenia.

[35] As noted in footnote 2, Stich and I also maintain that there are special purpose mechanisms for detecting one's own perceptual states. These too, are plausibly modular mechanisms of introspection.

[36] Both Peter Carruthers and Uta Frith have pressed this point (personal communication).

[37] For example Mark Ridley, *Evolution* (Cambridge, MA: Blackwell Science, 1993).

[38] See, for example, Todd Grantham and Shaun Nichols, "Evolutionary Psychology: Ultimate Explanations and Panglossian Predictions" in *Where Biology Meets Psychology : Philosophical Essays*, ed. V. Hardcastle. (Cambridge, MA: MIT Press, 1999); Shaun Nichols and Todd Grantham, "Adaptive Complexity and Phenomenal Consciousness," *Philosophy of Science* (forthcoming).

[39] For creatures that have language or some other sufficiently sensitive means of

communication, the Monitoring Mechanism would obviously be adaptive in *communicating* one's mental states to others. However, I don't want to assume that only creatures with sophisticated communication capacities evolved the Monitoring Mechanism.

[40] One line of evidence of egocentric mistakes in adults comes from work on the "false consensus effect". For a review, see Gary Marks and Norman Miller "Ten Years of Research on the False-Consensus Effect: An Empirical and Theoretical Review," *Psychological Bulletin*, 102 (1987): 72-90.

[41] See, for example, Nichols & Stich, *Mindreading.*

[42] For example, Michael Bratman, *Intentions, Plans, and Practical Reason* (Cambridge, MA: Harvard University Press, 1987); *Faces of Intention* (Cambridge: Cambridge University Press, 1999); Martha Pollack, "Overloading Intentions for Efficient Practical Reasoning," *Nous* 25 (1991): 513-536; "The Uses of Plans," *Artificial Intelligence* 57 (1992): 43-68.

[43] M. Bratman, D. Israel, and M. Pollack, "Plans and Resource-Bounded Practical Reasoning" in *Philosophy and AI: Essays at the Interface* ed. R. Cummins and J. Pollock (Cambridge, MA: Bradford/MIT Press, 1991).

[44] Pollack, "Overloading Intentions" and "The Uses of Plans".

[45] As will be evident, I am concerned with concepts as individuated by cognitive content, e.g., by the functional role of the concept. I want to remain neutral on the semantic issues surrounding concepts (see, e.g., Jerry Fodor, *Concepts: Where Cognitive Science Went Wrong* [Oxford: Oxford University Press, 1998]), so for present purposes, I want to leave open whether the cognitive content contributes to the semantic content.

[46] For a review, see Ziva Kunda, *Social Cognition* (Cambridge, MA: MIT Press, 1999).

[47] Gordon Gallup, "Chimpanzees: Self-recognition," *Science*, 167 (1970): 86-87; Michael

Lewis and Jeanne Brooks-Gunn, *Social Cognition and the Acquisition of the Self* (New York: Plenum Press, 1979).

[48] For example, Wellman, *The Child's Theory of Mind*; D. Hart, S. Fegley, Y. Chan, D. Mulvey, L. Fischer, "Judgments about Personal Identity in Childhood and Adolescence," *Social Development*, 2 (1993): 66-81.

[49] Gallup, "Chimpanzees: Self-recognition".

[50] Lewis & Brooks-Gunn, *Social Cognition and the Acquisition of the Self.*

[51] For a review of the comparative findings, see Michael Tomasello and Josep Call, *Primate Cognition* (Oxford: Oxford University Press, 1997).

[52] For example, Gallup, "Self-awareness and the Emergence of Mind in Primates," *American Journal of Primatology*, 2 (1982): 237-248.

[53] Daniel Povinelli, "Can Animals Empathize? Maybe Not," *Scientific American Presents* 9 (1998), 74; see also Povinelli, "The Unduplicated Self," in *The Self in Early Infancy*, ed. P. Rochat (Amsterdam: North-Holland/Elsevier, 1995), 161-192; Daniel Povinelli and Christopher Prince, "When Self Met Other," in *Self-awareness: Its Nature and Development* ed. M. Ferrari, R. Sternberg (New York : Guilford Press, 1998), 37-107.

[54] Povinelli and Prince, "When Self Met Other", 53.

[55] Ibid., 51.

[56] A similar line of reasoning suggests that it's possible for an agent to be fluent with personal pronouns without the agent having a psychological understanding of the self. Michael Lewis makes much of the fact that the toddler's use of personal pronouns emerges at the same time as mirror self-recognition, and he maintains that these both show understanding of the self (see, e.g., Michael Lewis and Douglas Ramsay, "Intentions, Consciousness, and Pretend Play," in

Developing Theories of Intention :Social Understanding and Self-control, ed. P. Zelazo, J. Astington, and D. Olson (Mahwah, N.J. : L. Erlbaum Associates, 1999), 89, 83). But, one might maintain that just because the young child uses the pronoun "I" does not show that the child has a concept of himself as a psychological subject. Indeed, using personal pronouns seems not to require that the subject have any understanding of the mind at all. It might merely indicate that the subjects understand something like the Body's "I".

The philosophical literature on self-conception inspired by Hector-Neri Casteñada (e.g., Castañeda, "'He': A Study in the Logic of Self-Consciousness" *Ratio,* 8 (1966): 130-157; John Perry, *The Problem of the Essential Indexical: And Other Essays* (Oxford: Oxford University Press, 1993); Georges Rey, *Contemporary Philosophy of Mind* (Oxford: Blackwell, 1997)) also tends to neglect this distinction between the concept of self as body and a concept of self as mind. For instance, in Perry's famous example, he recounts following a trail of sugar throughout the supermarket until he realizes that **he** is the one who is spilling the sugar. While this realization does implicate *some* concept of self, it does not require any understanding of the self-as-mind.

[57] N. Breen, D. Caine, M. Coltheart, J. Hendy, and C. Roberts, "Towards an Understanding of Delusions of Misidentification: Four Case Studies" *Mind and Language*, 15 (2000), 74-110.

[58] For example, Roy D'Andrade, "The Folk Model of the Mind" in *Cultural Models in Language and Thought* ed. D. Holland and N. Quinn (Cambridge: Cambridge University Press, 1987); *The Development of Cognitive Anthropology* (Cambridge: Cambridge University Press, 1995); Richard Nisbett and Lee Ross, *Human Inference* (Hillsdale, NJ: Lawrence Erlbaum, 1980); Wellman, *The Child's Theory of Mind*; Hart et al., "Judgments about Personal Identity in Childhood and Adolescence".

[59] Wellman, *The Child's Theory of Mind*, 297, 298.

[60] For example, Carol Guardo and Janis Bohan, "Development of a Sense of Self-Identity in Children," *Child Development*, 42 (1971): 1909-1921; Don Mohr, "Development of Attributes of Personal Identity," *Developmental Psychology*, 14 (1978): 427-428; Robert Selman, *The Growth of Interpersonal Understanding : Developmental and Clinical Analyses* (New York: Academic Press, 1980).

[61] Mohr "Development of Attributes of Personal Identity," 428.

[62] For example, Selman, *The Growth of Interpersonal Understanding*; Raymond Montemayor and Marvin Eisen "The Development of Self-conceptions from Childhood to Adolescence," *Developmental Psychology*, 13 (1977): 314-319.

[63] Hart et al. "Judgments about Personal Identity in Childhood and Adolescence".

[64] Ibid., 69.

[65] Ibid., 77-8.

[66] Despite this useful experimental foray into the Theory of Mind's "I", there might be considerable variation in the Theory of Mind's concept of self across cultures. There has been a large literature charting differences in conceptions of the self across cultures (e.g., Joan Miller, "Culture and the Development of Everyday Social Explanation," *Journal of Personality and Social Psychology*, 46 (1984): 961-978; Daniel Hart and Susan Fegley, "Children's Self-awareness and Self-understanding in Cultural Context," in *The Conceptual Self in Context* ed. U. Neisser and D. Jobling (Cambridge: Cambridge University Press, 1997); Francis Hsu, "The Self in Cross-cultural Perspective," in *Culture and Self: Asian and Western Perspectives*, ed. A. Marsella, G. DeVos and F. Hsu (New York: Tavistock Publications, 1985), 24-55; Catherine Lutz, *Unnatural Emotions: Everyday Sentiments on a Micronesian Atoll and Their Challenge to*

Western Theory (Chicago: University of Chicago Press, 1988); Angeline Lillard, "Ethnopsychologies: Cultural Variations in Theory of Mind," *Psychological Bulletin*, 123 (1998): 3-30). This literature suggests that there may be profound cross-cultural differences in self-conception. For instance, in her review of cross-cultural work on Theory of Mind, Angeline Lillard maintains that, while Europeans identify the self with the mind, there is evidence that the Japanese are less inclined to do so. "There is no clear, single division between mind and body, and self is not solely identified with mind" (12). Much of the cross-cultural data is difficult to interpret, however, since the data relies on older methodologies for assessing concept of self. Nonetheless, it's possible that there are radical cultural differences in the concept of self, as there seem to be radical cultural differences in other domains (see, e.g., R. Nisbett, K. Peng, I. Choi, and A. Norenzayan, "Culture and Systems of Thought: Holistic vs. Analytic Cognition" (forthcoming); Jonathan Weinberg, Shaun Nichols, and Stephen Stich, "Normativity and Epistemic Intuitions," (forthcoming)).

[67] Recall that the introspection mechanism is a mechanism that produces output *when activated*. And it's not clear that looking in a mirror would activate the mechanism. It may well be that animals have the *capacity* to access their own mental states without it being the case that they are prompted to do so when looking in the mirror.

[68] See note 66.

[69] Cf. Galen Strawson, "The Sense of Self" in *From Soul to Self*, ed. M. Crabbe (London: Routledge, 1999).

[70] It's also possible that this concept of self supports judgments about unity of self across mental states types. So, the Monitoring Mechanism produces the belief that *I believe that I'm typing* and the belief that *I want a drink*. Perhaps, then, the Monitoring Mechanism delivers the belief

that ***I*** believe that p and desire that q. However, there is an important complication here. For most purposes of this paper, it doesn't matter too much whether the Monitoring Mechanism is a single mechanism or several different mechanisms, one for each mental state type. However, in the present context, this might make a great deal of difference. If there is a single Monitoring Mechanism that generates self-attributions for the different types of mental states (beliefs, desires, imaginings, intentions), then it's easy to claim that this mechanism delivers a concept of self that is unified across mental state types. If, however, there are different Monitoring Mechanisms for different mental state types, then more needs to be said about whether the concept of self that is delivered is unified across mental state types. It's possible that the different mechanisms *Belief Monitoring Mechanism, Desire Monitoring Mechanism, Intention Monitoring Mechanism,* all produce outputs that have been designed to use the same representation "I". It's also possible that they produce different self-concepts. It then becomes an interesting empirical question when and how these concepts get unified.

[71] Reported in Bartsch and Wellman, *Children Talk about the Mind,* 41, 90, 52.

[72] Gopnik and Slaughter, "Young Children's Understanding of Changes in Their Mental States," 106.

[73] There are a number of interesting questions about the relation between the concept of self delivered by the Monitoring Mechanism and the concept of self delivered by Theory of Mind. One possibility is that the concept of self delivered by Theory of Mind emerges entirely independently from the concept of self delivered by the Monitoring Mechanism. It's also possible that the Theory of Mind largely builds and elaborates on the concept of self delivered by the Monitoring Mechanism. These issues deserve serious consideration, but they cannot be taken up here.

[74] Colin McGinn, *Problems in Philosophy: The Limits of Inquiry* (Oxford: Blackwell, 1993); this view is also adopted by Steven Pinker, *How the Mind Works* (New York: Norton, 1997).

[75] Bernard Williams, "The Self and the Future," *Philosophical Review*, 79 (1970): 161-180, 179; Thomas Nagel, "Subjective/Objective," in *Mortal Questions* (Cambridge: Cambridge University Press, 1979), 200-1; Simon Blackburn, "Has Kant Refuted Parfit?" in *Reading Parfit* ed. J. Dancy (Oxford: Blackwell, 1997), 180-201, 180-1.

[76] Thomas Nagel, "Subjective/Objective," 200.

[77] The distinction between 1st and 3rd person perspectives here is not, of course, fully marked by the pronouns that are used. We can, for instance, think of ourselves using Theory of Mind. And we can think of others by putting ourselves in their place (see e.g., Paul Harris, "From Simulation to Folk Psychology: The Case for Development," *Mind and Language* 7 (1992): 120-144; Stephen Stich and Shaun Nichols, "Cognitive Penetrability, Rationality, and Restricted Simulation," *Mind & Language*, 12 (1997): 297-326).

[78] Bernard Williams, "The Self and the Future".

[79] John Locke, *An Essay Concerning Human Understanding*, 2nd edition, 1694; Derek Parfit, *Reasons and Persons* (Oxford: Oxford University Press, 1984); Sydney Shoemaker, "Personal Identity: A Materialist's Account," In Shoemaker and Swinburne, *Personal Identity* (Oxford: Basil Blackwell, 1984); Williams, "The Self and the Future".

[80] See, for example, Locke's *Essay*, Shoemaker, "Personal Identity: A Materialist's Account," Robert Nozick, *Philosophical Explanations* (Cambridge, MA: Harvard University Press, 1981). Although Williams' 3rd person thought experiment is particularly effective at eliciting the intuition that psychological properties are crucial, Williams himself resists this conclusion since he thinks that the same situation elicits a very different intuition when considered from the 1st

person perspective.

[81] Thomas Nagel, *The View from Nowhere* (Oxford: Oxford University Press, 1986), 33.

[82] Interestingly, Nagel actually rejects a similar line of response:

> "The concept of 'someone' is not a generalization of the concept of 'I'. Neither can exist without the other, and neither is prior to the other. To possess the concept of a subject of consciousness an individual must be able in certain circumstances to identify himself and the states he is in without external observation. But these identifications must correspond by and large to those that can be made on the basis of external observation, both by others and by the individual himself" (Nagel, *The View from Nowhere*, 35)

So, according to Nagel, 'I' can't exist without 'someone' and neither is prior. However, in light of the cognitive account of introspection developed earlier in this paper, Nagel's claim is quite dubious. There is good reason to think that the Monitoring Mechanism & Theory of Mind are separate mechanisms, so it seems possible that the concept of 'I' can exist without the concept of 'someone'. Indeed, if, as seems likely, the Monitoring Mechanism comes on line earlier than Theory of Mind, the concept of 'I' is ontogenetically prior to the concept of 'someone'. And it's possible that one concept can be isolated under certain conditions, e.g., clever thought experiments.