Forthcoming in J. Prinz (ed.) *The Oxford Handbook on Philosophy of Psychology.* (New York: Oxford University Press)

Mindreading and the Philosophy of Mind[*]

Shaun Nichols

One key pitfall in the philosophy of psychology is to use psychology as window dressing on what is effectively an exercise in *a priori* philosophy. The other major pitfall is to use philosophy as window dressing on what is effectively a review of the scientific literature. The latter is the greater danger when it comes to "mindreading" – the everyday capacity to attribute mental states – because there is so much excellent science on mindreading. In an attempt to avoid these pitfalls, I will organize this essay around how research on mindreading might inform central questions in the philosophy of mind. To this end, I will begin with a review of some traditional concerns in philosophy of mind, and proceed to examine whether the psychological work helps address these concerns.[1]

1. Mindreading: The Philosophical Backstory

1.1. Traditional answers

Do mental states exist? How do we come to know about them in ourselves? How do we come to know about them in others? These are central questions in the philosophy of mind, but in early modern philosophy, only the last was a genuine problem. As for the other two questions, the answers were obvious and uncontroversial. We come to know about our own current mental states through introspection, an unimpeachable source. In the *Meditations*, the reader is led to doubt everything *except* what is happening in his own mind. Our access to our own current mental states is indubitable, and hence we can be sure that these mental states exist. Indeed, for centuries, philosophical orthodoxy allowed that we do have indubitable access to our own mental states. This also, of course, cements the metaphysical claim – mental states do exist.

The presumption of special access to one's own mental states makes the problem of *other* minds especially salient. There are actually multiple problems of other minds.

[*] Many thanks to Mike Bishop, Joshua Knobe, and Eric Schwitzgebel for comments on a previous draft.

[1] One issue that I will largely bypass is the debate between "theory theory" and "simulation theory". In its undiluted form, the theory theory maintains that we can explain mindreading by appealing only to information-bases dedicated to the psychological domain. Simulation theory in its undiluted form maintains that we can explain all of mindreading without appealing to any psychology-specific information bases; rather, mindreading is entirely driven by imagining oneself in the target's situation and then attributing the mental state that you find yourself in, in the imagination. There are two reasons I'm not reviewing this debate. First, the debate has increasingly moved to common ground – there are few if any advocates for either undiluted position; rather, most recent accounts of mindreading allow a place for both psychology-specific information bases and for simulation-like processes. Second, there are already a number of overviews of this work. For long treatments, see Currie & Ravenscroft 2002, Nichols & Stich 2003, and Goldman 2006. For shorter treatments, see Cruz & Goldman 2003; Stich & Nichols 2003; and Goldman & Mason 2007.

One familiar problem is the skeptical worry, how can I be certain that there are other minds? This problem will largely be set aside in this chapter. But philosophers stretching back to Augustine have been interested in a descriptive question, "how in fact do we come to have beliefs in other minds?" as well as a corresponding normative question, "are we justified in these beliefs?" These will be the problems about other minds explored here.

The descriptive problem: why do we believe in other minds?

Part of the traditional problem of other minds, from Augustine to the present, has been to ascertain the basis for our belief in other minds. Augustine gives what is perhaps the most familiar answer to this question – our belief in other minds is based on drawing an analogy from one's own mind. Thus, Augustine writes:

> when a living body is moved there is no way open to our eyes to see the mind, a thing which cannot be seen by the eyes. But we perceive something present in that mass such as is present in us to move our mass in a similar way; it is life and soul… Therefore we know the mind of anyone at all from our own; and from our own case we believe in that which we do not know. For not only do we perceive a mind, but we even know what one is, by considering our own (*De Trinitate* 8.6.9).

Mill's treatment is more familiar, also appealing to the role of analogy:

> By what evidence do I know, or by what considerations am I led to believe, that there exist other sentient creatures; that the walking and speaking figures which I see and hear, have sensations and thoughts, or in other words, possess Minds?…. I conclude that other human beings have feelings like me, because, first, they have bodies like me, which I know, in my own case, to be the antecedent condition of feelings; and because, secondly, they exhibit the acts, and other outward signs, which in my own case I know by experience to be caused by feelings (Mill 1865).

I recognize similarities between my body/behavior and the bodies/behaviors of others. Since I know that certain behaviors of mine are caused by mental states, when I see others exhibit similar behaviors, I reason by analogy to the conclusion that their behaviors are also caused by mental states.

Mill is sometimes interpreted differently, as proposing that our beliefs in other minds are inferred as the best explanation of the behavior we see (e.g. Thomas 2001). This view has been explicitly developed in recent years by a number of people, but the best known treatment comes from Robert Pargetter, who writes:

> What is the nature of the inferences that we all so commonly, and rightly, make from certain behavioural evidence to the mental lives of other people? This paper explores … the thesis that these inferences should best be viewed as being common scientific or hypothetic inferences, or arguments to the best explanation (1984, 158).

The suggestion here seems to be that we come to believe in other minds by using good inductive techniques – appealing to other minds is the best explanation for the behavior we observe.

A third answer to the descriptive question comes from Thomas Reid. According to Reid, the attribution of mental states doesn't depend on any kind of reasoning:

> As soon as children are capable of asking a question…they must be convinced, that those with whom they have this intercourse are intelligent beings….It is

> evident they are capable of such intercourse long before they can reason. Every one knows, that there is a social intercourse between the nurse and the child before it is a year old…Now I would ask how a child of a year old come by this conviction? Not by reasoning surely, for children do not reason at that age. Nor is it by external senses, for life and intelligence are not objects of the external senses… (1969[1785], 633)

Rather than analogy or any other form of reasoning, Reid takes the belief in other minds to be a "first principle", given by God:

> the belief of life and intelligence in other men, is absolutely necessary for us before we are capable of reasoning; and therefore the Author of our being hath given us this belief antecedently to all reasoning (635).

Reid is not entirely clear what it is to be a "first principle". But at a minimum, first principles are not generated by reasoning, either of the analogical or best explanation variety.

The justificatory problem: are we justified in believing in other minds?

The second problem concerns whether our beliefs about other minds are justified. This will depend, of course, on how we arrive at the beliefs and on what is required for justification. For now it will suffice to note that advocates for each of the above positions maintain that the way we *do* come to believe in other minds is in fact entirely epistemically proper. So we will rejoin the justificatory issue after we get a better sense of the right answer to the descriptive question. But before proceeding to the data, we need to pause for a revolution.

1.2. The Sellarsian revolution

In the twentieth century two developments led to a revolutionary new picture of lay views of the mind. The first development was the challenge to introspection as a source of knowledge about the mind. This challenge occurred in both psychology and philosophy. At the beginning of the 20th century, psychologists became suspicious about the reliability of introspection for several reasons, but perhaps the best known factor was an apparently intractable dispute between introspectionist psychologists in Germany and the U.S. If the introspectionists couldn't even agree on the basic introspective data, then this can hardly be a reliable method for learning about the mind.[2] Methodological behaviorism emerged, largely displacing introspectionism in mainstream scientific psychology. In effect, the methodological behaviorists forswore all talk of mental states in scientific psychology. In philosophy, a related attack was brewing, sometimes labeled "logical behaviorism", and typically associated with Ryle (1949) and Wittgenstein (1958). According to logical behaviorism, it is a mistake to think that there are beliefs and desires inhering in an unobservable mind. Unlike the methodological behaviorists, though, logical behaviorists did not enjoin against using terms like "belief" and "desire". Rather, logical behaviorists maintained that such terms refer not to internal mental states, but to publicly observable phenomena, in particular, to dispositions to behave in certain ways under certain conditions. Of course, one important consequence of logical

[2] Hurlburt (1993) argues that the disagreement between the schools was actually a disagreement about the *interpretation* of the introspective reports, not about the raw descriptions given by the subjects.

behaviorism is that since beliefs and desires are not internal states, they cannot be possibly be revealed by introspection. Even after behaviorism fell out of favor, skepticism about introspection continued to exercise a powerful hold over psychology and philosophy of mind.

The second historic development occurred in the wake of introspection's decline. If we can't rely on introspection to provide us with knowledge of the mind, then we need a new account of the source of our knowledge about the mind. Wilfrid Sellars (1956) developed what has turned out to be the most influential alternative to the introspectionist account of lay knowledge about the mind. Rather than maintain that the mind reveals its secrets to itself through introspection, Sellars suggested that lay people have a *theory* of the mind. The way Sellars makes the point is by proposing a myth about the origins of our commonsense view. He suggests that in our distant past, our ancestors never spoke of internal mental states like beliefs and desires. Rather, these "Rylean" ancestors only spoke of publicly observable phenomena like behavior and dispositions to behave. At this point, our ancestors even lacked terms for inner mental states. Then one day Jones, a great genius, arose from this group. Jones recognized that positing inner states like *thoughts* as theoretical entities provides a powerful basis for explaining the verbal behavior of his peers, and Jones developed a *theory* according to which such behavior is indeed the expression of internal thoughts. Jones then taught his peers how to use the theory to interpret the behavior of others. We are ultimately the beneficiaries of Jones' genius as well, since we too use the theory to interpret others' behavior.

Although Sellars explicitly presents this origin story as a myth, the point is that the myth allows us to see quite clearly a new picture of the nature of our commonsense views about the mind. On this picture, the commonsense view is a *theory* of mind, and the theory posits inner mental states like thoughts that are not publicly observable. Once this proposal is in place, the myth can be seen as one possible (and surely mistaken) account of the origin of the theory. The important advance is that Sellars has provided us with an alternative account of commonsense psychology that does not rely on introspection. Nor, however, does it adopt the desperate logical behaviorist account that terms like *thought* refer to publicly observable phenomena. This idea that folk psychology is a theory, whatever its origins, has come to be known as the "theory theory" (Morton 1980). The theory theory not only makes clear a new way to construe commonsense psychology once introspection had been displaced, it also provides a new way of construing introspection itself. For Sellars maintains that our ability to attribute mental states to ourselves arises out of the theory we use to attribute mental states to others.

This shift puts a surprising twist on the problem of other minds. The traditional problem starts from our knowledge of our own mind, and then asks how we can know about other minds. On a Sellarsian account, there is no such clean break. Our epistemic access to other minds is not much more problematic than our epistemic access to our own minds. In both cases, our access is mediated by a commonsense psychological theory.

1.3. The eliminativist threat

Thus far the philosophical issues have been cast as broadly epistemic. But if the Sellarsian account is right, it unleashes the specter of a radical new metaphysical view. For if lay views about the mind derive not from indubitable introspection but from a commonsense theory, then it may well be the case that lay views of the mind will not cohere with mature scientific views of the mind. It is at this point that the most prominent theme in contemporary philosophical discussions of mindreading emerges. For, some suggest, if the folk account diverges widely from the scientific account, then we should conclude that the folk theory is *wrong*. Indeed, it may turn out that the folk theory is so thoroughly wrong that we must reject the theoretical posits of "belief" and "desire" and acknowledge that beliefs and desires don't really exist. According to Eliminative Materialism this is exactly the case. The folk theory is so far off the mark that we need to extirpate the ontology of folk psychology entirely, just as we have extirpated the ontology of the supernatural.

Eliminativist arguments have been developed in two rather different ways. Some (e.g. Stich 1983) maintain that folk psychology will be at odds with a mature scientific psychology and that this gives reason to suspect that we need to jettison the folk ontology of beliefs and desires. Others (e.g. Churchland 1981) envision neuroscience as the proper scientific approach to the mind and argue that folk psychology is a hopelessly mistaken theory that will not fit with mature neuroscience; as a result, the folk ontology should be rejected in favor of a neuroscientific ontology.[3]

It is tempting to respond to the eliminativist by claiming that introspection directly reveals that mental states exist. However, in good Sellarsian fashion, Churchland replies that the introspection-based defense "makes the same mistake that an ancient or medieval person would be making if he insisted that he could just see with his own eyes that the heavens form a turning sphere, or that witches exist." For, Churchland continues:

> all observation occurs within some system of concepts, and our observation judgments are only as good as the conceptual framework in which they are expressed. In all three cases--the starry sphere, witches, and the familiar mental states-- precisely what is challenged is the integrity of the background conceptual frameworks in which the observation judgments are expressed. To insist on the validity of one's experiences, traditionally interpreted, is therefore to beg the very question at issue. For in all three cases, the question is whether we should reconceive the nature of some familiar observational domain (1988, 47-8).

In the present case, folk psychology is the flawed background framework and according to Churchland this flawed framework contaminates introspection just as much as it contaminates the attribution of mental states to others.

Thus, in the 20th century the philosophical landscape was rototilled beyond recognition. In early modern philosophy, we could be sure of our own mental states, but less sure whether others have mental states. Now philosophers tell us that science will reveal that there aren't any mental states.

2. Mindreading: One's own mind

With the history behind us, let's move on to the evidence. Many cognitive scientists in fact still maintain that the available evidence indicates that the only way to

[3] Evaluating Eliminative Materialism is a matter of considerable complexity. For, as several authors have noted, eliminativist arguments typically depend on important assumptions about reference, reductionism, and other controversial issues in metaphysics (see, e.g., Lycan 1988, Stich 1996, Mallon et al. forthcoming). I'll set aside these concerns here.

access one’s own mental states is via the theory of mind (e.g., Gopnik 1993, Carruthers 1996). This view effectively embraces Sellars’ early suggestion that the folk theory of mind is essential not only for attributing mental states to others but also for attributing mental states to oneself. Although this approach to introspection continues to be extremely influential, there are empirical reasons to think that, as revolutions often do, this one went too far.

The Sellarsian view maintains that the folk psychological theory we use to attribute mental states to others is essential to self attribution as well. One way to attack such a thesis is to show a dissociation. If we find individuals who systematically fail at other-attribution, but succeed at self attribution, this will pose a prima facie problem for the Sellarsian account.

According to the Sellarsian view, self-attribution depends on the general folk psychological theory. Hence, where the folk psychological theory is defective, this should affect both other attribution and self attribution. Gopnik & Meltzoff 1994 maintain that 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 indicates otherwise. For on a wide range of tasks, children do not exhibit the expected parallel performance (see Nichols & Stich 2003). Children are capable of attributing knowledge and ignorance to themselves before they are capable of attributing those states to others (Wimmer et al. 1988); they are capable of attributing certain perceptual states to themselves before they are capable of attributing such states to others (Gopnik & Slaughter 1991); 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 (German & Leslie forthcoming).

Data on autism provide another way to explore possible dissociations. Autistic children are widely regarded as having deficient mindreading capacities. For instance, autistic children tend to fail to appreciate false beliefs in others (Baron-Cohen et al.1985). And studies of spontaneous speech indicates that autistic children basically never talk about cognitive mental states like beliefs or thoughts (Tager-Flusberg 1993). Given this deficit in understanding other minds, the Sellarsian predicts that autistic children should be similarly impaired at detecting their own mental states. However, there is some intriguing evidence suggesting that autistic children do have access to their current thoughts. In a recent set of studies, Farrant and colleagues found that autistic children did remarkably well on “metamemory” tests (Farrant et al. 1999). 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. 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.” (Farrant et al. 1999, 108) 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, children were asked to remember a set of numbers. 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."

The foregoing suggests that self-attribution can be intact even when the folk psychological theory is apparently defective. Further evidence against the Sellarsian account emerges from the developmental data when one considers the *kinds* of mistakes that toddlers make about other minds (Nichols 2000). When asked what another person 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". 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." (Gopnik & Meltzoff 1997, 116). 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. These experimental findings of egocentric desire attributions are corroborated by ecological reports. 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 (Hoffman 1982). This is a puzzle for the Sellarsian. For if the child does not yet understand how to attribute desires and perceptions to others, then the Sellarsian predicts that she should also be incapable of drawing on her own desires and perceptions to make such attributions. Egocentric mistakes indicate an asynchrony – the young child apparently exploits information about her own mental states to guide attribution before she has the relevant theory of others' mental states.[4]

The foregoing evidence suggests that we need alternative cognitive accounts of introspection according to which we have some access to our own minds that does not depend on the theory of mind. Stephen Stich and I (2003) draw on this kind of evidence to argue 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 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.* 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 was that the Monitoring Mechanism is an independent introspection mechanism for detecting a large class of one's own mental states. This

[4] 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 the crucial piece of theory is, 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.

model, of course, is a rejection of the Sellarsian idea that our access to our own minds must be mediated by a general theory of mind. (See Goldman 1993, 2006 for alternative anti-Sellarsian approaches.)

Let's now return to the metaphysical challenge thrown by eliminative materialism. The revival of introspection reestablishes a kind of Cartesian response to eliminativism. As we saw, one tempting reply to eliminativism is to say that we know that mental states exist from our own introspections. But, drawing on the Sellarsian framework, Churchland had a ready reply: introspection is laden by the very folk psychological theory that is being called into question. Now, however, it seems that this argument doesn't reach far enough to secure whole-scale eliminativism. Even if our folk theory is wrong, our introspection can be immune to these polluting influences. For if the anti-Sellarsian models of introspection are right, for a large class of mental states, our introspective access is not routed through our folk psychology. While introspection might well depend on some kind of conceptual structures, they are not given by the folk psychological theory that has been brought under attack by eliminativists.[5]

Before we break out the bourbon, though, we need to acknowledge an important qualification on this reply to eliminativism. The Sellarsian revolution went too far. But for many of our self-attributions, the Sellarsian account is probably right. Consider, for instance, the attribution of the emotion *schadenfreude*. It's unlikely that one can detect schadenfreude in oneself before one has the chunk of folk-psychological theory about what the emotion is. Moreover, it's likely that much self-attribution is theory-laden in this way. It's helpful at this point to distinguish between mental state attributions that are laden with folk psychological theory and those that aren't. We can call the former "thick" mental state attributions and the latter "thin" (Nichols & Stich 2003, 205-209). The monitoring mechanisms provide *thin* mental state attributions that are insulated from the eliminativist arguments. We get information about our own minds through a channel that is not subject to the potentially distorting influence of folk psychological theory. But we also make *thick* mental state attributions to ourselves, and these *are* subject to the potentially distorting influence of folk psychological theory.

Hence, the monitoring mechanism account provides a kind of Cartesian response to the eliminativist. At least part of introspection is not laden by folk-psychological theory. So the (alleged) deficiencies of that theory provide no argument that the deliverances of introspection are themselves defective. But this introspective reply only goes so far. Since thick attributions are infused with folk psychological theory, introspection provides no special defense against eliminativism concerning thick mental states.

[5] However, even if the Monitoring Mechanism account is right, some might maintain that the mental state attributions that are generated by the mechanisms do not involve the concepts *belief, desire,* etc. For, the objection goes, those concepts implicate a rich set of inferential connections, and the outputs of the Monitoring Mechanisms are inferentially impoverished (Nichols & Stich 2003, 162-3). This issue depends on unresolved issues about concept individuation, but if the objection holds, we can simply maintain that the outputs of the Monitoring Mechanism exploit "proto-concepts" of *belief, desire,* etc. (Nichols & Stich 2003, 163). In that case, one can claim that these attributions of proto-beliefs, proto-desires, etc., are not threatened by the eliminativist arguments.

3. Mindreading: Other minds

3.1. The descriptive problem

One useful way to carve up the domain of mindreading is to distinguish between the processes involved in attributing a mind at all, and the processes involved in attributing an elaborate set of mental states to something that is already coded as a mind. Let's call the first category *agency attribution* and the second *mental state elaboration.* We'll begin by looking at the mechanisms underlying agency attributions and consider which philosophical account is the best fit. Then we will turn to mental state elaboration and follow the same drill. Finally, we will consider whether the mechanisms produce justified beliefs.

Agency attribution

As it happens, there are several different paths by which we come to attribute minds. Here, I'll focus on mechanisms that are known to emerge early in development.

One of the earliest, and most enjoyable, entries on agency attribution is Heider & Simmel's study on how adults describe an animation involving geometric objects (1944). They found overwhelmingly that adults describe the animation by adverting to mental states. For example, almost everyone says that the big triangle and little triangle fight at one point in the animation, and that the big triangle tries to get out of the box at another point. More recent work shows that children respond in much the same way to these sorts of stimuli. Like the adults in Heider & Simmel's study, when children are asked to describe what they saw, they advert to the goals, beliefs, and intentions of triangles in a 2D animation (e.g. Bowler & Thommen 2000; Abell et al. 2000; Dasser 1989). If you've watched one of these, the results will come as no surprise. It's extremely natural, indeed compelling, to see these objects as minded. Why is that? Because the motion trajectories of the triangles push the right buttons to trigger agency attribution. This becomes evident when one contrasts Heider-Simmel style animations with an animation of triangles moving about the surface in straight lines at constant speeds. In that case, there is no inclination to start attributing mental states to the triangles. Motion isn't enough. But it remains possible that quite simple motion cues suffice for agency attributions. For instance, change in speed plus change in direction might be sufficient to generate an attribution of mind, even if nothing can be discerned about the goals or thoughts of that mind (Scholl & Tremoulet 2000, 305). Of course, as adults, we don't cave to our first-blush intuitions of agency here – we know, on slight reflection, that the images aren't agents. Nonetheless, there clearly is a mechanism that generates these powerful, if overridable, intuitions of agency, and this mechanism likely plays an important role in everyday attributions of agency.

Susan Johnson and colleagues use much different techniques to discern mechanisms underlying agency attribution. There are several ways that a baby might reveal that she thinks an object has a mind: she might follow the 'gaze' of the object, try to communicate with the object, imitate the behaviors of the object, or attribute goals to the object. By exploiting this variety of indicators, Johnson provides strong evidence that infants attribute agency as a result of particular kinds of cues. In one experiment, 12 month old infants were shown a fuzzy brown object under a variety of different conditions (Johnson, Slaughter & Carey 1998). In one condition, the fuzzy brown object (with no facial features) interacted contingently with the infant by beeping and flashing when the infant babbled or moved; in another condition, the fuzzy brown object exhibited an equivalent amount of flashing and beeping, but in this condition the activity was not contingent on the infant's behavior. In both conditions, children's looking behavior was measured when the fuzzy brown object 'gazed' at one of two objects by making a smooth, 45 degree turn towards the object and remaining in this orientation for several seconds. What Johnson and colleagues found was that infants would follow the "gaze" of the fuzzy brown object, but only when its beeping and flashing was contingent. In another set of conditions the fuzzy brown object didn't flash or beep, but Johnson and colleagues found that babies were more likely to follow the 'gaze' of the fuzzy brown object when it had eyes than when it didn't. In other experiments, babies were shown a stuffed orangutan that had a face and exhibited contingent interaction. Babies imitated the behavior of the stuffed animal and made communicative gestures to it, indicating that it coded the object as an agent (Johnson et al. 2001). In more recent experiments, Johnson and colleagues devised a new object, the blob, a bright green object about the shape of an adult shoe that had no facial features but could beep and move around on its own. In one condition, a confederate engages the blob in "small talk", and the blob beeped contingently with the confederate; in the other condition the blob beeped randomly. Again they found that babies would follow the 'gaze' of the blob, but only in the contingent interaction condition. Finally, and perhaps most impressively, the blob design has recently been coupled with Amanda Woodward's (1998) goal attribution experiment. Woodward (1998) showed babies an arm moving towards one of two objects. Then the locations of the objects were switched. Babies looked longer when the arm reached to the same location that now had a different object, suggesting that they expected the person to reach for the same goal-object. Shimizu & Johnson (2004) found something similar with the blob – babies looked longer when the blob moved in the same direction but towards a different object. But this *only* occurred when the blob had behaved contingently with the confederate.[6]

Insects provide an intriguing real-world test case here. For insects exhibit all of the cues reviewed above – eyes, distinctive motion trajectories, and contingent interaction. So, do children attribute mental states to insects? It would seem so. In a Japanese study, subjects were asked "Does a grasshopper feel something if a person who has been taking care of it daily dies?" Although most Japanese adults denied that the grasshopper would feel anything, about ¾ of the six year olds maintained that the grasshopper would feel something, e.g., "the grasshopper will feel unhappy" (Inagaki & Hatano 1991, 225).

For each of the cues we've reviewed, there is presumably a mechanism (or set of mechanisms) that transitions from the perceptual cues to the intuitive attribution of agency. There are, of course, other ways to come to attribute agency (e.g., by testimony from other), but we will be primarily concerned with the cluster we've reviewed here, so let's call these the "agency attribution mechanisms". All of these mechanisms seem to be

[6] One very interesting question that has not yet been explored is whether babies would attribute *conscious states* to the blob. For instance, in the contingent interaction condition, if the blob was poked with a pin, would babies expect the blob to feel pain? I suspect that they would, and if that turns out to be the case, it would support the view that we naturally attribute consciousness to creatures coded as agents.

preserved into adulthood. Heider & Simmel's studies were, of course, initially conducted on adults. And Johnson found that adults described the fuzzy brown object's turning behavior in mentalistic terms in the same conditions that generated gaze-following in infants (Johnson et al. 1998, p. 237).

Now that we've charted several early emerging paths to agency attribution, let's turn to the descriptive accounts reviewed in section 1.1. Are these processes analogical? Presumably not. As for attributing agency from physiognomic triggers this doesn't seem to be an instance of analogical reasoning. That is, the infant doesn't think "Hey, I've got eyes and a mind; that thing has eyes; so it probably has a mind." In the contingent behavior experiments, there's no close analogy between the child's body/behavior and the body/beeping of the blob. Obviously there's even less of an analogy between a child and a triangle in a 2D animation. So the analogical account is a poor fit for describing these fundamental tracks to attributing agents.

Perhaps the processes involved in the agency attribution mechanisms more closely resemble inference to the best explanation. But at least for the physiognomic triggers in Johnson's experiments, the inference to the best explanation account is clearly wrong. Surely the best explanation for why a fuzzy brown object has eyes is not "because it has mental states". For the animation and contingent interaction cases, the situation is somewhat more complicated. But in both cases, before any motion occurs, adults fully know that the triangles and the blob don't have minds. But this doesn't eradicate the inclination to attribute agency. So the mechanisms that drive these attributions of agency seem to be insensitive to the kind of global considerations that characterize inference to the best explanation.

Finally, what about Reid's suggestion that agency attribution is based on a divinely given *first principle*? I'm not so sure about the divine part of the equation, but given the early onset of agency attribution, and the rather stubborn character of the processes, it does seem plausible that these processes are built-in. As Reid anticipated, our early beliefs in other minds are not achieved by reasoning to that conclusion (either by analogy or by inference to the best explanation). Rather, there is a hodgepodge of relatively simple mechanisms that generate these intuitions of agency.

Mental state elaboration

When an object has been credited with a mind, the child doesn't attribute consciousness or intentionality in the abstract. That is, the child doesn't think "that thing is conscious" or "that creature has thoughts *about things*". Rather, the child attributes particular conscious or intentional mental states. How does this elaboration of mental state attribution proceed?

As with agency attribution, mental state attribution is achieved through a variety of pathways. Under some conditions, even young children will use surprisingly sophisticated techniques for discerning what another person's discrepant beliefs and desires are. For instance, children use facial expressions to infer likes & dislikes (Repacholi & Gopnik 1997); children use searching behavior to infer false beliefs (Bartsch & Wellman 1989), children use line of sight information to infer that a person lacks perceptual access to an object (Flavell 1978). And by the time we're adults, at least some of us sometimes deploy more broad-based reasoning, something akin to inference to the best explanation. This is illustrated, for instance, in the efforts of our colleagues in the history of philosophy. They use all the available resources to best determine what Descartes thought about, say, innateness. But people also exploit processes that are much less deliberative, and much less sensitive to differences between themselves and the object of interest. Children and adults engage in *egocentric* attributions, and this is probably the single most fecund source of our everyday expectations about others' mental states. Since this kind of attribution is so pervasive, and since it resonates with the venerable analogical argument, I will focus entirely on this kind of mental state elaboration.

Egocentric attribution as default

In *Mindreading*, Stich and I maintained that in the absence of any distinguishing evidence, adults attribute beliefs by relying on a default strategy of attributing their own beliefs to others (Nichols & Stich 2003, 66-7, 85). The same is true of children, of course. Furthermore, children also use an egocentric strategy to attribute desires. When children are not given information about another person's preferences, they tend to think that the other will have preferences like their own (Astington & Gopnik 1991, 44; Martin et al. 1995; Repacholi & Gopnik 1997).[7]

The most systematic investigation of egocentric attributions comes from social psychology. In the "assumed similarity" (Cronbach 1955) and "attributive projection" (Holmes 1968) paradigms, the subject is asked to note her own view along a continuous dimension and then to rate the views of another individual. This literature produced ample demonstration that one's own mental states affect one's attributions of mental states to others (see Marks & Miller 1987 for a review). In the last thirty years, researchers have turned to a different methodology for exploring egocentric attribution. In this methodology, developed by Lee Ross and his colleagues (Ross et al. 1977), subjects estimate the degree of consensus for a given claim or decision; Ross and colleagues argued that subjects exhibit a "false consensus effect" – they assume that their own views are more representative of the population than they really are. In a recent experiment in this tradition, Krueger & Clement (1994) presented subjects with forty statements from a personality questionnaire (the MMPI-2). After the presentation of each item, the subjects were asked to note whether they agreed or disagreed with the statement. In a later portion of the experiment, the same forty statements were presented to subjects again, and for each statement, subjects were told to "enter the percentages between 0 and 100 that best reflect your belief about the proportion of people who would agree with each statement" (598). By looking at subjects' responses across multiple items, Krueger & Clement show that subjects systematically over-project their own judgments (Krueger & Clement 1994).

[7] One might argue that this is the product of children thinking that what they desire is objectively good, and then expecting others to desire the good as well. However, children seem not to be uniformly objectivist about their preferences. For instance, children tend to say that watermelon is yummy, but when asked whether it's "yummy for real" or "yummy for some people", children tend to say "yummy for some people". The situation is exactly the opposite for moral properties – hitting is "bad for real" (Nichols & Folds-Bennett 2003). Despite their recognition that the *yumminess* of watermelon depends on the individual, the earlier results suggest that children probably egocentrically attribute their watermelon preference to others.

The false consensus effect emerges not only for beliefs and desires, but also for feelings. Krueger & Clement found the false consensus effect for statements such as:

I certainly feel useless at times.
Criticism or scolding hurts me terribly.
I am so touchy on some subjects that I can't talk about them (1994, 600).

Even for these statements, subjects who endorsed a given statement estimated a much higher percentage of agreement for that statement than did non-endorsers. This indicates that one's own emotional experience affects one's attribution of emotion to others. Indeed, earlier work in social psychology suggests that one's own current emotional state affects one's judgment about another's current emotional state. Singer & Feshbach (1962) found that subjects with exam-induced anxiety were more likely than control subjects to attribute anxiety to individuals shown in pictures. And Feshbach & Singer (1957) found that subjects who had been frightened by electrical shocks characterized another college aged male as more fearful than did subjects who hadn't been exposed to electrical shock.

The work in social psychology brings out several important points about egocentric attribution. First, the false consensus effect isn't diminished even if subjects are explicitly told about the bias before they make their estimations (Krueger & Clement 1994). As a result, Krueger & Clement refer to the false consensus effect as an "ineradicable" egocentric bias. Second, even when subjects are given information about the responses of others, this has only a slight impact on the egocentric effect (Clement & Krueger 2000; Alicke & Largo 1995; Krueger & Clement 1994). Third, there is some experimental reason to suspect that the egocentric attribution is the default response. People use it as an initial strategy under conditions of uncertainty, but they discard it when there is overriding information about the target; in addition, people show an even more pronounced egocentric bias when they're under time pressure (Epley et al. 2004). And finally, the egocentric bias is mitigated when subjects are asked to make judgments about the views of people in out-groups (Robbins & Krueger 2005), even when the groups are arbitrarily defined. For example, in one task, subjects completed a task that was supposed to determine their "cognitive style". Subjects were randomly assigned to classifications of "Figurer" or "Grounder". Then subjects were asked to make their consensus estimates either about others with the same cognitive style or about those with the alternate cognitive style. Subjects showed a stronger egocentric bias when giving consensus estimates for their in-group than for the out-group (Clement & Krueger 2002).

Egocentric attributions and atypical populations

Another important source of evidence on egocentrism comes from atypical populations. For instance, if we look to the attributional capacities of individuals with affective abnormalities, we may find that there is a direct link between emotional capacity and attributional capacity.

In fact, there *are* subjects who seem to be impaired in emotion attribution in precisely those areas in which they are emotionally deficient. Psychopaths are generally thought to lack the capacity for guilt. Indeed, lack of guilt is one of the diagnostic criteria for identifying psychopaths (Hare 1991). Psychopaths seem to show a corresponding deficiency in their understanding of guilt. In his book on psychopathy, Robert Hare (1993) presents serial killer Ted Bundy's discussion of guilt as representative of the psychopath's conception of guilt. In an interview from death row, Bundy said:

> Guilt? It's this mechanism that we use to control people. It's an illusion. It's a kind of social control mechanism - and it's *very* unhealthy. It does terrible things to our bodies. And there are much better ways to control our behavior than that rather extraordinary use of guilt.
>
> It doesn't solve anything, necessarily. It's just a very gross technique we impose upon ourselves to control the people, groups of people. I guess I am in the enviable position of not having to deal with guilt. There's just no reason for it (reported in Michaud & Aynesworth 1989, 288).

Bundy seems not to have a grasp of the internal, involuntary aspect of guilt. Hare gives another example that shows another presumed psychopath's deficient conception of guilt: "When asked if he experienced remorse over a murder he'd committed, one young inmate told us, 'Yeah, sure, I feel remorse.' Pressed further, he said that he didn't "feel bad inside about it" (Hare 1993, 41).

These statements from presumed psychopaths are striking, and they certainly suggest that psychopaths have a deficient understanding of guilt. Moreover, recent research on psychopaths indicates a general deficiency in the psychopath's capacity to attribute guilt. James Blair and his colleagues (Blair et al. 1995) found that psychopaths are perfectly adept at attributing sadness and happiness to others. Psychopaths even perform well in attributing the social emotion of embarrassment. However, psychopaths perform poorly in attributing guilt. In the experiment, they presented psychopaths and control criminals with stories and then asked the subjects how the protagonist in the story would feel. In one of the stories, one boy punches another boy. After being presented with this story, the subjects were asked, "How do you think that person [referring to the boy who punched] would feel in that situation?". Control criminals tended to say that the puncher would feel guilty; psychopaths tended to say that the puncher would feel happy! There are delicate issues about what can be concluded from this study. But the tempting conclusion is that the understanding of emotions is incomplete where emotional experience is incomplete (see Goldman 2006 for reviews of related evidence).

Perhaps it's not all that surprising that we need to draw on our own experiences to understand such experiences in others. It would be remarkable, for instance, if children suffering from congenital insensitivity to pain (CIP) had a normal appreciation of pain in others. Indeed, in *Life without Pain*, a recent documentary on children with CIP, there's a surreal scene in which the mother of such a child describes how she tried to communicate the notion of pain to her daughter by first pushing a needle deep into her daughter's arm and then explaining that she wouldn't be able to do the same to herself because it would hurt.

So, the foregoing suggests that knowledge of other minds depends crucially on one's own experiences. Whether surprising or not, this again marks a major break with Sellarsian approaches. For it seems that we *have* to draw on our own experience to understand the minds of others. Furthermore, it seems that our inclination to engage in such egocentric attribution is rooted in innate components of mindreading. Children don't *learn* to be egocentric. Rather, that's where they begin.

With some details of egocentric attributions in hand, we can turn to the philosophical accounts. Does egocentric attribution count as inference to the best explanation? Obviously not. It is not used in a statistically sensitive way – even when people are given explicit evidence that ought to overturn their egocentric bias, they still show the bias (e.g., Clement & Krueger, 2000; Alicke & Largo, 1995; Krueger & Clement, 1994). This tendency to egocentric attribution is not something that is well integrated into a broad theory. Indeed, much egocentric attribution, like the default attribution of belief, seems not to be explanation at all.

Can the pattern of egocentric attribution be assimilated to the argument from analogy? If we construe analogy along classical lines (a la Augustine and Mill) as tied to similarities in behavior between self and other, then egocentric attribution far outstrips analogy. For egocentric attribution often occurs in the absence of any observed behavior. For instance, people are not monitoring behavior when they make their judgments in false consensus experiments. But obviously, on a less uptight interpretation of analogy, egocentric attribution fits quite well. Like analogy, it involves projecting one's own mental states on others; and also like analogy, it is based on a presumption of similarity, as reflected in the fact that the egocentric effect diminishes when dissimilarities are registered (as for out-groups). So the spirit of the analogical account seems to be vindicated in egocentric attributions, which comprise a huge sector of our mental state attributions.

3.2. Justificatory problem

Now that we know something about how our beliefs about other minds are triggered and elaborated, we can turn to the normative question: are those beliefs justified? Of course, to do this, we need to have a clearer idea about what is involved in justification. Let's consider here two of the most prominent accounts of justification: internalism and reliabilism. Roughly speaking, internalists claim that a belief can only be justified if the justification is internally accessible to the epistemic agent. Take Billy's belief that Susie is in pain. In order for Billy's belief to be justified, Billy must have cognitive access to the justificational support for the belief. Reliabilists, on the other hand, allow that some of the justificational support for a belief can be external to the epistemic agent. According to the most familiar version of the view (Goldman 1979, 1986), what matters particularly for justification is that the process that led to the belief is a reliable process, i.e., one that tends to produce a high proportion of true beliefs.

Let's begin with the basic phenomenon of agency attribution. As set out above, there are several simple mechanisms that lead people to believe that an object has a mind. By internalist scruples, do these beliefs count as justified? Well, one might worry about the fact that we seem to make what we later realize to be mistakes in the attribution of other minds. Indeed, in the studies reviewed above, *all* of the attributions of minds are mistaken – the triangles, the fuzzy brown object, and the blob don't have minds. In each case, the experimenter triggers an attribution of mind where there is none. Reid already recognizes the problem: "It cannot be said, that the judgments we form concerning life and intelligence in other beings are at first free from error." Although Reid recognizes that people make mistakes in the attribution of minds, he didn't find the errors worrying: "the errors of children in this matter lie on the safe side; they are prone to attribute intelligence to things inanimate. These errors are of small consequence, and are gradually corrected by experience and ripe judgment" (1969[1785], 635)

Despite Reid's insouciance, for internalists, the justificatory problem that emerges is serious. People's confidence in the consciousness of other creatures is produced by a mechanism that is known to produce lots of false positives. So the mechanisms themselves do not come with any cognitively accessible justification. Whatever the process is, it's not very epistemically sensitive. Thus, if we want internalist justifications for the attributions issued by the agency attribution mechanisms, they must come from some other source.

The problem isn't limited to false positives; a complementary problem emerges with the possibility of false negatives when we have intuitions that an object or collection is not a conscious agent. Consider Ned Block's (1978) notorious thought experiment in which we are told to imagine that the inhabitants of China all follow a script designed so that the nation of China will be executing the functions that are operative in one person's mind. Most of us have a strong intuition that the nation, so described, is not conscious. But how seriously should we take that intuition? In light of the evidence on agency attribution, we need to acknowledge that part of the explanation for our intuition might be that Block's description of the system doesn't push our AGENCY buttons – no physiognomy, no contingent interaction, no distinctive motion cues.[8] A parallel problem arises even for the brain. As many philosophers have emphasized (e.g. McGinn 1989), when we think about the brain as a massive collection of neurons with various functions and physical characteristics, it seems bizarre that this unwieldy amalgam is conscious. But again, we are considering a description that steers clear of the AGENCY buttons. Those buttons, as we've seen, involve fairly stupid systems that are sensitive to fairly specific sorts of cues, and in light of this, we can hardly expect the resulting intuitions to identify the presence or absence of genuine consciousness. Thus, when we focus on our justification for having the intuitions of agency, it seems that the mechanism is too crude to generate beliefs that bring along cognitively accessible justifications.

From a reliabilist perspective, the justificatory situation might seem somewhat less dire. For here we needn't ask whether we can tell from the inside whether our intuitions are justified. Rather, we simply ask whether the intuitions are produced by a mechanism that tends to produce a high ratio of true beliefs. Even here the justificatory situation is problematic, though. For the mechanisms of agency attribution can easily be triggered by false positives, and this might directly undercut their reliability. There is, of course, an obvious explanation for why agency attribution mechanisms have hair triggers – false positives are less costly than false negatives. This gives prudential reasons for favoring the kinds of mechanisms we have, but it doesn't directly alleviate the epistemic problem. It's good that the agency attribution mechanisms err on the side of caution, but that does little to establish their justificatory credentials. Of course, it all depends on how often the mechanisms generate false beliefs, but it's far from obvious that the mistakes are so infrequent that the reliabilist needn't worry about them.

Thus, the agency attribution mechanisms do not give us an easy escape from the justificatory problem of other minds. There remain serious concerns, both for internalists and reliabilists, about whether those mechanisms produce justified beliefs. Let's turn now to the elaboration of mental states. And let's focus on cases that are not plagued by the

[8] This is not intended as a defense of functional accounts of consciousness; rather, the point is to highlight the difficulties with relying on the intuitions delivered by our agency attribution mechanisms.

general justificatory worries about agency attribution. We'll just assume that there are lots of other minds, including all normally functioning mammals. Are the egocentric attributions that we make to those minds justified?

The pattern of egocentric attribution, as we have seen, has important similarities with analogical arguments. Psychologists have often characterized these attributions as the product of infantile mistakes (e.g., Piaget) or the desire to feel part of the majority (e.g., Holmes 1968). Philosophers have been even less accommodating. Philosophers have long lampooned the analogical argument as being statistically preposterous. The familiar criticism is that drawing a conclusion about other minds on the basis of a single case is terrible inductive practice. Hence, if we believe in other minds because of analogy, that belief is in bad epistemic repair. Here is a recent presentation of the criticism from that pulse of our philosophical times, the Stanford Encyclopedia of Philosophy:

> the analogical arguer's own experience is crucial to the analogical inference. This becomes the target of the classical and ongoing objection to this inference; that it is a generalisation based on one case only and therefore fatally unsound (e.g., Malcolm, 1962, p. 152). This feature is seen as so problematic that the one element common to all other responses to the problem of other minds is a desire to avoid having our own experience play the central role in the evidence (Hyslop 2005; see also Churchland 1979, 90).

If this objection is right, it would clearly extend to the kinds of egocentric attributions that we have found to be pandemic to mindreading. That is, if the criticism is right, there is no way we would be justified in our egocentric attributions. There have been attempts to shore up the analogical argument (Hill 1984; Hyslop 1995). But these accounts are too intellectualist to enhance the justificatory credentials of the largely unreflective kinds of egocentric attributions that we actually find.

On an internalist approach, it seems hard to avoid the conclusion that typical egocentric attribution is not justified. One's own feelings and thoughts might – *really* might – be idiosyncratic. The lay person is unable to tell, drawing only on what is cognitively accessible to her, whether her mental states are idiosyncratic or not. Hence, her egocentric attributions aren't justified. Her own mental state provides her with a statistically negligible sample, hence she lacks introspectively available information to justify her practice of egocentric attribution.

On a reliabilist account, the situation is very different indeed. For a reliabilist, the statistical complaint is in fact statistically naïve. To evaluate whether egocentric attribution is reliable, we do not restrict attention only to what is introspectively accessible. Rather, we need to know about the distribution of attributed traits. To see this, let's consider the reliability of the process of egocentrically attributing a mental state to someone poked by a pin. On the egocentric attribution story, if a person sees an object (previously coded as an agent) poked with a pin, he will attribute pain to the agent just in case he believes that being poked by a pin would cause him to feel pain. Now, is this a reliable strategy? Well, let's take the apparent facts at face value:

(i) The vast majority of familiar agents do in fact typically feel pain in such cases.

(ii) A small minority, individuals with congenital insensitivity to pain, never feel pain in such cases.

Notice that, given the apparent facts, if everyone follows the egocentric strategy nearly all of the attributions will be accurate. We will be wrong in the tiny minority of cases in which the attributee has congenital insensitivity to pain. Children with congenital insensitivity to pain, of course, would be wrong about the vast majority of cases. But from the perspective of assessing global accuracy, that's a minimal cost. The egocentric strategy is highly reliable in these cases.

Of course, even if the above is right about pain attribution, it doesn't follow that the egocentric strategy is generally so reliable. Indeed, we saw in section 3.1 that people routinely make egocentric *mistakes* in the social psychology lab. But the fact that the egocentric attribution strategy is accident prone in social psychology labs does little to impugn its epistemic status as a reliable strategy. While the false consensus task has taught us a great deal about egocentric attribution, the task itself is hardly representative of day-to-day mindreading. If the fast and frugal heuristics program has taught us anything, it's that sometimes the erroneous responses elicited in social psychology experiments are driven by simple mechanisms that work very well indeed in the real world. It's quite plausible that this is the case with egocentric attribution. Somewhat surprisingly, then, despite the abuse piled on the argument from analogy, our analogy-based beliefs might have rather good epistemic credentials after all.

References:

Abell, F., Happé, F. and Frith, U. (2000). Do Triangles Play Tricks? Attribution of Mental States to Animated Shapes in Normal and Abnormal Development. *Journal of Cognitive Development* **15**: 1-20.

Alicke, M. and Largo, E. (1995). The Unique Role of the Self in the False Consensus Effect. *Journal of Experimental Social Psychology* **31**: 28-47.

Astington, J. and Gopnik, A. (1991). Understanding Desire and Intention. In *Natural Theories of Mind*, ed. A. Whiten. Cambridge: Basil Blackwell.

Baron Cohen, S., A.M. Leslie and U. Frith (1985). Does the Autistic Child Have a 'Theory of Mind'? *Cognition* **21**: 37-46.

Bartsch, K. and Wellman, H. (1989). Young Children's Attribution of Action to Belief and Desires. *Child Development* **60**: 946-64.

Blair, R., Sellers, C., Strickland, I., Clark, F., Williams, A., Smith, M. and Jones, L. (1995). Emotion Attributions in the Psychopath. *Personality and Individual Differences* **19**: 431-437.

Block, N. (1978). Troubles with Functionalism. In *Perception and Cognition: Issues in the Foundations of Psychology*, ed. W. Savage. Minneapolis: University of Minnesota Press.

Bowler, D. and Thommen, E. (2000). Attribution of Mechanical and Social Causality to Animated Displays by Children with Autism. *Autism* **4**: 147–71.

Carruthers, P. (1996). Autism as Mind-Blindness: An Elaboration and Partial Defence. In *Theories of Theories of Mind*, ed. P. Carruthers and P. Smith. Cambridge: Cambridge University Press.

Churchland, P. (1981). Eliminative Materialism and the Propositional Attitudes. *Journal of Philosophy* **LXXVII**(2): 67-90.

Churchland, P. M. (1979). *Scientific Realism and the Plasticity of Mind*: Cambridge

University Press.

Churchland, P. M. (1988). *Matter and Consciousness*: MIT Press.

Clement, R. and Krueger, J. (2000). The Primacy of Self-Referent Information in Perceptions of Social Consensus. *British Journal of Social Psychology* **39**: 279-299.

Clement, R. and Krueger, J. (2002). Social Categorization Moderates Social Projection. *Journal of Experimental Social Psychology* **38**: 219-231.

Cronbach, L. (1955). Processes Affecting Scores on "Understanding of Others" and "Assumed Similarity." *Psychological Bulletin* **52**: 177-93.

Cruz, J. and Gordon, R. M. (2003). Simulation Theory. In *The Encyclopedia of Cognitive Science*. London: Nature Publishing Group.

Currie, G. and Ravenscroft, I. (2002). *Recreative Minds: Imagination in Philosophy and Psychology*: Oxford University Press.

Dasser, V., Ulbaek, I. and Premack, D. (1989). The Perception of Intention. *Science* **243**: 365-7.

Epley, N., Keysar, B., Van Boven, L. and Gilovich, T. (2004). Perspective Taking as Egocentric Anchoring and Adjustment. *Journal of Personality & Social Psychology* **87**: 327-39.

Farrant, A., Boucher, J. and Blades, M. (1999). Metamemory in Children with Autism. *Child Development* **70**: 107-31.

Feshbach, S. and Singer, R. (1957). The Effects of Fear Arousal and Suppression of Fear Upon Social Perception. *Journal of Abnormal and Social Psychology* **55**: 283-288.

Flavell, J. (1978). The Development of Knowledge About Visual Perception. In *Nebraska Symposium on Motivation, V. 25*, ed. C. Keasey. Lincoln: University of Nebraska Press.**:** 43-76.

German, T. P. and Leslie, A. (forthcoming). Self-Other Differences in False Belief: Recall Versus Reconstruction.

Goldman, A. (1979). What Is Justified Belief? In *Justification and Knowledge*, ed. G. Pappas. Dordrecht: Reidel**:** 1-23.

Goldman, A. (1986). *Epistemology and Cognition*. Cambridge, MA: Harvard University Press.

Goldman, A. (1993). The Psychology of Folk Psychology. *Behavioural and Brain Sciences* **16**: 15-28.

Goldman, A. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. New York: Oxford University Press.

Goldman, A. and Mason, K. (2007). Simulation. In *Philosophy of Psychology and Cognitive Science*, ed. P. Thagard. Elsevier.

Gopnik, A. (1993). How We Know Our Own Minds: The Illusion of First-Person Knowledge of Intentionality. *Behavioral and Brain Sciences* **16**: 1-14.

Gopnik, A. and Meltzoff, A. (1994). Minds, Bodies, and Persons: Young Children's Understanding of the Self and Others as Reflected in Imitation and Theory of Mind Research. In *Self-Awareness in Animals and Humans*, ed. S. Parker, R. Mitchell and M. Boccia. New York: Cambridge University Press.

Gopnik, A. and Meltzoff, A. (1997). *Words, Thoughts, and Theories*. Cambridge, MA: MIT Press.

Gopnik, A. and Slaughter, V. (1991). Young Children’s Understanding of Changes in Their Mental States. *Child Development* **62**: 98-110.

Hare, R. (1991). *The Hare-Psychopathy Checklist-Revised*. Toronto: Multi-Health Systems.

Hare, R. (1993). *Without Conscience: The Disturbing World of the Psychopaths among Us*. New York: Pocket Books.

Heider, F. and Simmel, M. (1944). An Experimental Study of Apparent Behavior. *American Journal of Psychology* **57**: 243-259.

Hill, C. (1984). In Defense of Type Materialism. *Synthese* **59**: 295-320.

Hoffman, M. (1982). Development of Prosocial Motivation: Empathy and Guilt. In *Development of Prosocial Behavior*, ed. N. Eisenberg. New York: Academic Press**:** 281-313.

Holmes, D. (1968). Dimensions of Projection. *Psychological Bulletin* **69**: 248-68.

Hurlburt, R. (1993). *Sampling Inner Experience in Disturbed Affect*. New York: Plenum Press.

Hyslop, A. (1995). *Other Minds*. Dordrecht: Kluwer.

Hyslop, A. (2005). Other Minds. *The Stanford Encyclopedia of Philosophy (Winter 2005 Edition)*, ed. E. Zalta. From http://plato.stanford.edu/archives/win2005/entries/other-minds/.

Inagaki, K. and Hatano, G. (1991). Constrained Person Analogy in Young Children’s Biological Inference. *Cognitive development* **6**: 219-31.

Johnson, S., Booth, A. and O’Hearn, K. (2001). Inferring the Goals of Non-Human Agents. *Cognitive Development* **16**: 637–656.

Johnson, S., Slaughter, V. and Carey, S. (1998). Whose Gaze Will Infants Follow? Features That Elicit Gaze-Following in 12-Month-Olds. *Developmental Science* **1**: 233-238.

Krueger, J. and Clement, R. (1994). The Truly False Consensus Effect: An Ineradicable and Egocentric Bias in Social Perception. *Journal of Personality and Social Psychology* **67**: 596-610.

Lycan, W. (1988). *Judgement and Justification*. Cambridge: Cambridge University Press.

Mallon, R., Machery, E., Nichols, S., and Stich, S. (forthcoming). Against Arguments from Reference.

Marks, G. and Miller, N. (1987). Ten Years of Research on the False-Consensus Effect: An Empirical and Theoretical Review. *Psychological Bulletin* **102**: 72-90.

Martin, C., Eisenbud, L. and Rose, H. (1995). Children's Gender-Based Reasoning About Toys. *Child Development* **66**: 1453-71.

McGinn, C. (1989). Can We Solve the Mind-Body Problem? *Mind* **98**: 349-366.

Michaud, S. and Aynesworth, H. (1989). *Ted Bundy: Conversations with a Killer*. New York: New American Library.

Mill, J. (1865). *An Examination of Sir William Hamilton's Philosophy*. London: Longmans.

Morton, A. (1980). *Frames of Mind*: Oxford University Press.

Nichols, S. (2000). The Mind’s “I” and the Theory of Mind’s “I”: Introspection and Two Concepts of Self. *Philosophical Topics* **28**: 171-99.

Nichols, S. and Folds-Bennett, T. (2003). Are Children Moral Objectivists? Children’s Judgments About Moral and Response-Dependent Properties. *Cognition* **90**: B23-

B32.

Nichols, S. and Stich, S. (2003). *Mindreading. An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*: Oxford University Press.

Pargetter, R. (1984). The Scientific Inference to Other Minds. *Australasian Journal of Philosophy* **62**: 158-63.

Reid, T. (1969[1785]). *Essays on the Intellectual Powers of Man*. Cambridge, MA: MIT Press.

Repacholi, B. and Gopnik, A. (1997). Early Understanding of Desires: Evidence from 14 and 18 Month Olds. *Developmental Psychology* **33**: 12-21.

Robbins, J. and Krueger, J. (2005). Social Projection to Ingroups and Outgroups: A Review and Meta-Analysis. *Personality and Social Psychology Review* **9**: 32-47.

Ross, L., Greene, D. and House, P. (1977). The False Consensus Phenomenon: An Attributional Bias in Self-Perception and Social-Perception Processes. *Journal of Experimental Social Psychology* **13**: 279-301.

Ryle, G. (1949). *The Concept of Mind.* London: Hutchinson.

Scholl, B. and Tremoulet, P. (2000). Perceptual Causality and Animacy. *Trends in Cognitive Sciences* **4**: 299-309.

Sellars, W. (1956). Empiricism and the Philosophy of Mind. *Minnesota Studies in the Philosophy of Science* **1**: 253-329.

Shimizu, Y. and Johnson, S. (2004). Infants' Attribution of a Goal to a Morphologically Unfamiliar Agent. *Developmental Science* **7**: 425-30.

Singer, R. and Feshbach, S. (1962). Effects of Anxiety in Psychotics and Normals Upon the Perception of Anxiety in Others. *Journal of Personality* **30**: 574-587.

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

Stich, S. and Nichols, S. (2003). Folk Psychology. In *The Blackwell Guide to Philosophy of Mind*, ed. T. Warfield and S. Stich. Oxford: Blackwell.

Stich, S. P. (1983). *From Folk Psychology to Cognitive Science: The Case against Belief*: MIT Press. A Bradford Book.

Tager-Flusberg, H. (1993). What Language Reveals About the Understanding of Minds in Children with Autism. In *Understanding Other Minds: Perspectives from Autism*, ed. S. Baron-Cohen, H. Tager-Flusberg and D. Cohen. Oxford: Oxford University press**:** 138-57.

Thomas, J. (2001). Mill's Arguments for Other Minds. *British Journal for the History of Philosophy* **9**: 507-23.

Wimmer, H., Hogrefe, G. and Perner, J. (1988). Children's Understanding of Informational Access as a Source of Knowledge. *Child Development* **59**: 386-96.

Wittgenstein, L. (1958). *Philosophical Investigations*. Oxford: Blackwell.

Woodward, A. (1998). Infants Selectively Encode the Goal Object of an Actor's Reach. *Cognition* **69**: 1-34.