This paper was published in Stephen Stich & Ted A. Warfield, eds., *The Blackwell Guide to Philosophy of Mind,* (Oxford: Basil Blackwell) 2003. Pp. 235-255.

Folk Psychology

Stephen Stich
Rutgers University
&
Shaun Nichols
University of Utah

For the last 25 years discussions and debates about commonsense psychology (or "folk psychology," as it is often called) have been center stage in the philosophy of mind. There have been heated disagreements both about what folk psychology is and about how it is related to the scientific understanding of the mind/brain that is emerging in psychology and the neurosciences. In this chapter we will begin by explaining why folk psychology plays such an important role in the philosophy of mind. Doing that will require a quick look at a bit of the history of philosophical discussions about the mind. We'll then turn our attention to the lively contemporary discussions aimed at clarifying the philosophical role that folk psychology is expected to play and at using findings in the cognitive sciences to get a clearer understanding of the exact nature of folk psychology.

1. Why does folk psychology play an important role in the philosophy of mind?

To appreciate philosophers' fascination with folk psychology, it will be useful to begin with a brief reminder about the two most important questions in the philosophy of mind, and the problems engendered by what was for centuries the most influential answer to one of those questions. The questions are the Mind-Body Problem, which asks how mental phenomena are related to physical phenomena, and the Problem of Other Minds, which asks how we can know about the mental states of other people. On Descartes' proposed solution to the Mind-Body Problem, there are two quite different sorts of substances in the universe: physical substance, which is located in space and time, and mental substance, which is located in time but not in space. Mental phenomena, according to Descartes, are events or states occurring in a mental substance, while physical phenomena are events or states occurring in a physical substance. Descartes insisted that there is two-way causal interaction between the mental and the physical, though many philosophers find it puzzling how the two could

interact if one is in space and the other isn't. Another problem with the Cartesian view is that it seems to make the Other Minds Problem quite intractable. If, as Descartes believed, I am the only person who can experience my mental states, then there seems to be no way for you to rule out the hypothesis that I am a mindless Zombie – a physical body that merely behaves as though it was causally linked to a mind.

In the middle of the 20th century the verificationist account of meaning had a major impact on philosophical thought. According to the verificationists, the meaning of an empirical claim is closely linked to the observations that would verify the claim. Influenced by verificationism, philosophical behaviorists argued that the Cartesian account of the mind as the "ghost in the machine" (to use Ryle's memorable image) was profoundly mistaken. (Ryle, 1949) If ordinary mental state terms like 'belief', 'desire' and 'pain' are to be meaningful, they argued, they can't refer to unobservable events taking place inside a person (or, worse still, not located in space at all). Rather, the meaning of sentences invoking these terms must be analyzed in terms of conditional sentences specifying how someone would behave under various circumstances. So, for example, a philosophical behaviorist might suggest that the meaning of

(1) John believes that snow is white.

could be captured by something like the following:

(2) If you ask John, 'Is snow white' he will respond affirmatively.

Perhaps the most serious difficulty for philosophical behaviorists was that their meaning analyses typically turned out to be either obviously mistaken or circular – invoking one mental term in the analysis of another. So, for example, contrary to (2), even though John believes that snow is white, he may not respond affirmatively unless he is paying *attention*, *wants* to let you know what he thinks, *believes* that this can be done by responding affirmatively, etc.

While philosophical behaviorists were gradually becoming convinced that there is no way around this circularity problem, a very similar problem was confronting philosophers seeking verificationist accounts of the meaning of scientific terms. Verificationism requires that the meaning of a theoretical term must be specifiable in terms of observables. But when philosophers actually tried to provide such definitions, they always seemed to require additional theoretical terms. (Hempel, 1964) The reaction to this problem in the philosophy of science was to explore a quite different account of how theoretical terms get their meaning. Rather than being defined exclusively in terms of observables, this new account proposed, a cluster of theoretical

terms might get their meaning collectively by being embedded within an empirical theory. The meaning of any given theoretical term lies in its theory-specified interconnections with other terms, *both observational and theoretical*. Perhaps the most influential statement of this view is to be found in the work of David Lewis (1970, 1972). According to Lewis, the meaning of theoretical terms is given by what he calls a "functional definition." Theoretical entities are "defined as the occupants of the causal roles *specified by the theory*…; as *the* entities, whatever those may be, that bear certain causal relations *to one another* and to the referents of the O[bservational]-terms." (Lewis, 1972, p. 211; first & last emphasis added)

Building on an idea first suggested by Wilfrid Sellars (1956), Lewis went on to propose that ordinary terms for mental or psychological states could get their meaning in an entirely analogous way. If we "think of commonsense psychology as a term-introducing scientific theory, though one invented before there was any such institution as professional science," then the "functional definition" account of the meaning of theoretical terms in science can be applied straightforwardly to the mental state terms used in commonsense psychology. (Lewis, 1972, p. 212) And this, Lewis proposed, is the right way to think about commonsense psychology:

> Imagine our ancestors first speaking only of external things, stimuli, and responses …until some genius invented the theory of mental states, with its newly introduced T[heoretical] terms, to explain the regularities among stimuli and responses. But that did not happen. Our commonsense psychology was never a newly invented term-introducing scientific theory – not even of prehistoric folk-science. The story that mental terms were introduced as theoretical terms is a myth.
>
> It is, in fact, Sellars' myth …. And though it is a myth, it may be a good myth or a bad one. It is a good myth if our names of mental states do in fact mean just what they would mean if the myth were true. I adopt the working hypothesis that it is a good myth. (1972, 212-213)

In the three decades since Lewis and others[1] developed this account, it has become the most widely accepted view about the meaning of mental state terms. Since the account maintains that the meanings of mental state terms are given by functional definitions,

[1] Though we will focus on Lewis' influential exposition, many other philosophers developed similar views including Putnam (1960), Fodor & Chihara (1965), and Armstrong (1968).

the view is often known as *functionalism*.[2] We can now see one reason why philosophers of mind have been concerned to understand the exact nature of commonsense (or folk) psychology. According to functionalism, *folk psychology is the theory that gives ordinary mental state terms their meaning*.

A second reason for philosophers' preoccupation with folk psychology can be explained more quickly. The crucial point is that, according to accounts like Lewis', folk psychology is an *empirical* theory which is supposed to explain "the regularity between stimuli and responses" to be found in human (and perhaps animal) behavior. And, of course, if commonsense psychology is an empirical theory, it is possible that, like any empirical theory, it might turn out to be *mistaken*. We might discover that the states and processes intervening between stimuli and responses are not well described by the folk theory that fixes the meaning of mental state terms. The possibility that commonsense psychology *might* turn out to be mistaken is granted by just about everyone who takes functionalism seriously. However, for the last several decades a number of prominent philosophers of mind have been arguing that this is more than a *mere* possibility. Rather, they maintain, a growing body of theory and empirical findings in the cognitive and neurosciences strongly suggest that commonsense psychology *is* mistaken, and not just on small points. Rather, as Paul Churchland, an enthusiastic supporter of this view puts it:

> FP [folk psychology] suffers explanatory failures on an epic scale, …it has been stagnant for at least twenty-five centuries, and … its categories appear (so far) to be incommensurable with or orthogonal to the categories of the background physical sciences whose long term claim to explain human behavior seems undeniable. Any theory that meets this description must be allowed a serious candidate for outright elimination. (Churchland, 1981, 212)

Churchland does not stop at discarding (or "eliminating") folk psychological theory. He and other "eliminativists" have also suggested that, because folk psychology is such a seriously defective theory, we should also conclude that the theoretical terms embedded in folk psychology don't really refer to anything. Beliefs, desires and other posits of folk psychology, they argue, are entirely comparable to phlogiston, the ether, and other posits of empirical theories that turned out to be seriously mistaken; like phlogiston, the ether and the rest, *they do not exist*. Obviously, these are enormously provocative claims. Debating their plausibility has been high on the agenda of

---

[2] Though beware. In the philosophy of mind, the term 'functionalism' has been used for a variety of views. Some of them bear a clear family resemblance to the one we've just sketched while others do not. For good overviews see Lycan (1994) and Block (1994).

philosophers of mind every since they were first suggested.[3] Since the eliminativists' central thesis is that folk psychology is a massively mistaken theory, philosophers of mind concerned to evaluate that thesis will obviously need a clear and accurate account of what folk psychology is and what it claims.

2. What is folk psychology? Two possible answers

Functionalists, as we've seen, maintain that the meaning of ordinary mental state terms is determined by the role they play in a commonsense psychological theory. But what, exactly, is this theory? In the philosophical and cognitive science literature there are two quite different approaches to this question.[4] For Lewis, and for many of those who have followed his lead, commonsense or folk psychology is closely tied to the claims about mental states that almost everyone would agree with and take to be obvious.

> Collect all the platitudes you can think of regarding the causal relations of mental states, sensory stimuli, and motor responses…. Add also the platitudes to the effect that one mental state falls under another – 'toothache is a kind of pain' and the like. Perhaps there are platitudes of other forms as well. Include only platitudes that are common knowledge among us – everyone knows them, everyone knows that everyone else knows them, and so on. For the meanings of our words are common knowledge, and I am going to claim that *names of mental states derive their meaning from these platitudes*. (1972, p. 212 emphasis added)

So, on this approach, folk psychology is just a collection of platitudes, or perhaps, since that set of platitudes is bound to be large and ungainly, we might think of folk psychology as a set of generalizations that systematizes the platitudes in a perspicuous way. A systematization of that sort might also make it more natural to describe folk psychology as a theory. We'll call this the *platitude account* of folk psychology.

The second approach to answering the question focuses on a cluster of skills that have been of considerable interest to both philosophers and psychologists. In many cases people are remarkably good at *predicting* the behavior of other people. Asked to predict what a motorist will do as she approaches the red light, almost everyone says

---

[3] For an overview of these debates, see Stich (1996), Ch. 1, and the essay by ??? in this volume.

[4] The distinction was first noted in Stich & Ravenscroft (1994).

that she will stop, and fortunately our predictions are usually correct. We are also often remarkably good at *attributing* mental states to other people[5] – at saying what they perceive, think, believe, want, fear and so on, and at *predicting* future mental states and *explaining* behavior in terms of past mental states.[6] In recent discussions, the whimsical label *mindreading* has often been used for this cluster of skills, and during the last fifteen years developmental and cognitive psychologists have generated a large literature aimed at exploring the emergence of mindreading and explaining the cognitive mechanisms that underlie it.

The most widely accepted view about the cognitive mechanisms underlying mindreading (and until about a dozen years ago the *only* view), is that people have a rich body of mentally represented information about the mind, and that this information plays a central role in guiding the mental mechanisms that generate our attributions, predictions and explanations. Some of the psychologists who defend this view maintain that the information exploited in mindreading has much the same structure as a scientific theory, and that it is acquired, stored and used in much the same way that other commonsense and scientific theories are. These psychologists often refer to their view as *the theory theory*. (Gopnik & Wellman, 1994; Gopnik & Meltzoff, 1997) Others argue that much of the information utilized in mindreading is innate and is stored in mental "modules" where it can only interact in very limited ways with the information stored in other components of the mind. (Scholl & Leslie, 1999). Since modularity theorists and theory-theorists agree that mindreading depends on a rich body of information about how the mind works, we'll use the term *information-rich theories* as a label for both of them. These theories suggest another way to specify the theory that (if functionalists are right) fixes the meaning of mental state terms – it is the theory (or body of information) that underlies mindreading. We'll call this the *mindreading account* of folk psychology.

Let's ask, now, how the platitude account of folk psychology and the mindreading account are related. How is the mentally represented information about the mind posited by information-rich theories of mindreading related to the collection of platitudes that, according to Lewis, determines the meaning of mental state terms?

---

[5] Though not always, as we'll see in Section 4.

[6] Eliminativists, of course, would not agree that we do a good job at attributing and predicting mental states or at explaining behavior in terms of past mental states, since they maintain that the mental states we are attributing do not exist. But they would not deny that there is an impressive degree of *agreement* in what people say about other people's mental states, and that that agreement needs to be explained.

One possibility is that the platitudes (or some systematization of them) is near enough *identical* with the information that guides mindreading – that mindreading invokes little or no information about the mind beyond the commonsense information the everyone can readily agree to. If this were true then the platitude account of folk psychology and the mindreading account would converge. But, along with most cognitive scientists who have studied mindreading, we believe that this convergence is *very* unlikely. One reason for our skepticism is the comparison with other complex skills that cognitive scientists have explored. In just about every case, from face recognition (Young, 1998) to decision making (Gigerenzer et al., 1999) to commonsense physics (McCloskey, 1983; Hayes, 1985), it has been found that the mind uses information and principles that are simply not accessible to introspection. In these areas our minds use a great deal of information that people can not recognize or assent to in the way that one is supposed to recognize and assent to Lewisian platitudes. A second reason for our skepticism is that in many mindreading tasks people appear to attribute mental states on the basis of cues that they are not aware they are using. For example, Ekman has shown that there is a wide range of "deception cues" that lead us to believe that a target does not believe what he is saying. These include "a change in the expression on the face, a movement of the body, an inflection to the voice, a swallowing in the throat, a very deep or shallow breath, long pauses between words, a slip of the tongue, a micro facial expression, a gestural slip" (Ekman 1985, 43). In most cases, people quite unaware of the fact that they are using these cues. So, while there is still much to be learned about mental mechanisms underlying mindreading, we think it is very likely that the information about the mind that those mechanisms exploit is substantially richer than the information contained in Lewisian platitudes.

If we are right about this, then those who think that the functionalist account of the meaning of ordinary mental state terms is on the right track will have to confront a quite crucial question: Which account of folk psychology picks out the theory that actually determines the meaning of mental state terms? Is the meaning of these terms fixed by the theory we can articulate by collecting and systematizing platitudes, or is it fixed by the much richer theory that we can discover only by studying the sort of information exploited by the mechanisms underlying mindreading?

We don't think there is any really definitive answer to this question. It would, of course, be enormously useful if there were a well motivated and widely accepted general theory of meaning to which we might appeal. But, notoriously, there is no such theory. Meaning is a topic on which disagreements abound even about the most fundamental questions, and there are many philosophers who think that the entire functionalist approach to specifying the meaning of mental state terms is utterly

wrongheaded.[7] Having said all this, however, we are inclined to think that those who are sympathetic to the functionalist approach should prefer the mindreading account of folk psychology over the platitude account. For on the mindreading account, folk psychology is the theory that people actually use in recognizing and attributing mental states, in drawing inferences about mental states, and in generating predictions and explanations on the basis of mental state attributions. It is hard to see why someone who thinks, as functionalists do, that mental state terms get their meaning by being embedded in a theory would want to focus on the platitude-based theory whose principles people can easily acknowledge, rather than the richer theory that is actually guiding people when they think and talk about the mind.

3. The challenge from simulation theory

Let's take a moment to take stock of where we are. In Section 1 we explained why folk psychology has played such an important role in recent philosophy of mind: functionalists maintain that folk psychology is the theory that implicitly defines ordinary mental state terms, and eliminativists (who typically agree with functionalists about the meaning of mental state terms) argue that folk psychology is a seriously mistaken theory, and that both the theory and the mental states that it posits should be rejected. In Section 2, we distinguished two different accounts of folk psychology, and we argued, albeit tentatively, that functionalists should prefer the mindreading account on which folk psychology is the rich body of information or theory that underlies people's skill in attributing mental states and in predicting and explaining behavior. In this Section, we turn our attention to an important new challenge that has emerged to all of this. In the last dozen years a number of philosophers and psychologists have argued that it is a mistake to think that mindreading invokes a rich body of information about the mind. Rather, they maintain, mindreading can be explained as a kind of *mental simulation* that requires little or no information about how the mind works. (Gordon, 1986; Heal, 1986; Goldman, 1989; Harris, 1992) If these simulation theorists are right, and if we accept the mindreading account of folk psychology, then *there is no such thing as folk psychology*. That would be bad news for functionalists. It would also be bad news for eliminativists, since if there is no such thing as folk psychology then their core argument – which claims that folk psychology is a seriously mistaken theory – has gone seriously amiss.

---

[7] See, for example, Fodor & LePore (1992). For a useful overview of many of the disputes about the theory of meaning, see Devitt (1996).

How could it be that the mental mechanisms underlying mindreading do not require a rich body of information? Simulation theorists often begin their answer by using an analogy. Suppose you want to predict how a particular airplane will behave in certain wind conditions. One way to proceed would be to derive a prediction from aeronautical theory along with a detailed description of the plane. Another, quite different, strategy would be to build a model of the plane, put it in a wind tunnel that reproduces those wind conditions, and then simply observe how the model behaves. The second strategy, unlike the first, does not require a rich body of theory. Simulation theorists maintain that something like this second strategy can be used to explain people's mindreading skills. For if you are trying to predict what another person's mind will do, and if that person's mind is similar to yours, then you might be able to use components of your own mind as models of the similar components in the mind of the other person (whom we'll call the "target").

Here is a quick sketch of how the process might work. Suppose that you want to predict what the target will decide to do about some important matter. The target's mind, we'll assume, will make the decision by utilizing a decision making or "practical reasoning" system which takes his relevant beliefs and desires as input and (somehow or other) comes up with a decision about what to do. The lighter lines in Figure 1 are a sketch of the sort of cognitive architecture that might underlie the normal process of decision making. Now suppose that your mind can momentarily take your decision making system "off-line" so that you do not actually act on the decisions that it produces. Suppose further that in this off-line mode your mind can provide your decision making system with some hypothetical or "pretend" beliefs and desires – beliefs and desires that you may not actually have but that the target does. Your mind could then simply sit back and let your decision making system generate a decision. If your decision making system is similar to the target's, and if the hypothetical beliefs and desires that you've fed into the off-line system are close to the ones that the target has, then the decision that your decision making system generates will be similar or identical to the one that the target's decision making system will produce. If that off-line decision is now sent on to the part of your mind that generates predictions about what other people will do, you will predict that that's the decision the target will make, and there is a good chance that your prediction will be correct. All of this happens, according to simulation theorists, with little or no conscious awareness on your part. Moreover, and this of course is the crucial point, the process does not utilize any theory or rich body of information about how the decision making system works. Rather, you have simply used your own decision making system to *simulate* the decision that the target will actually make. The dark lines in Figure 1 sketch the sort of cognitive architecture that might underlie this kind of simulation-based prediction.

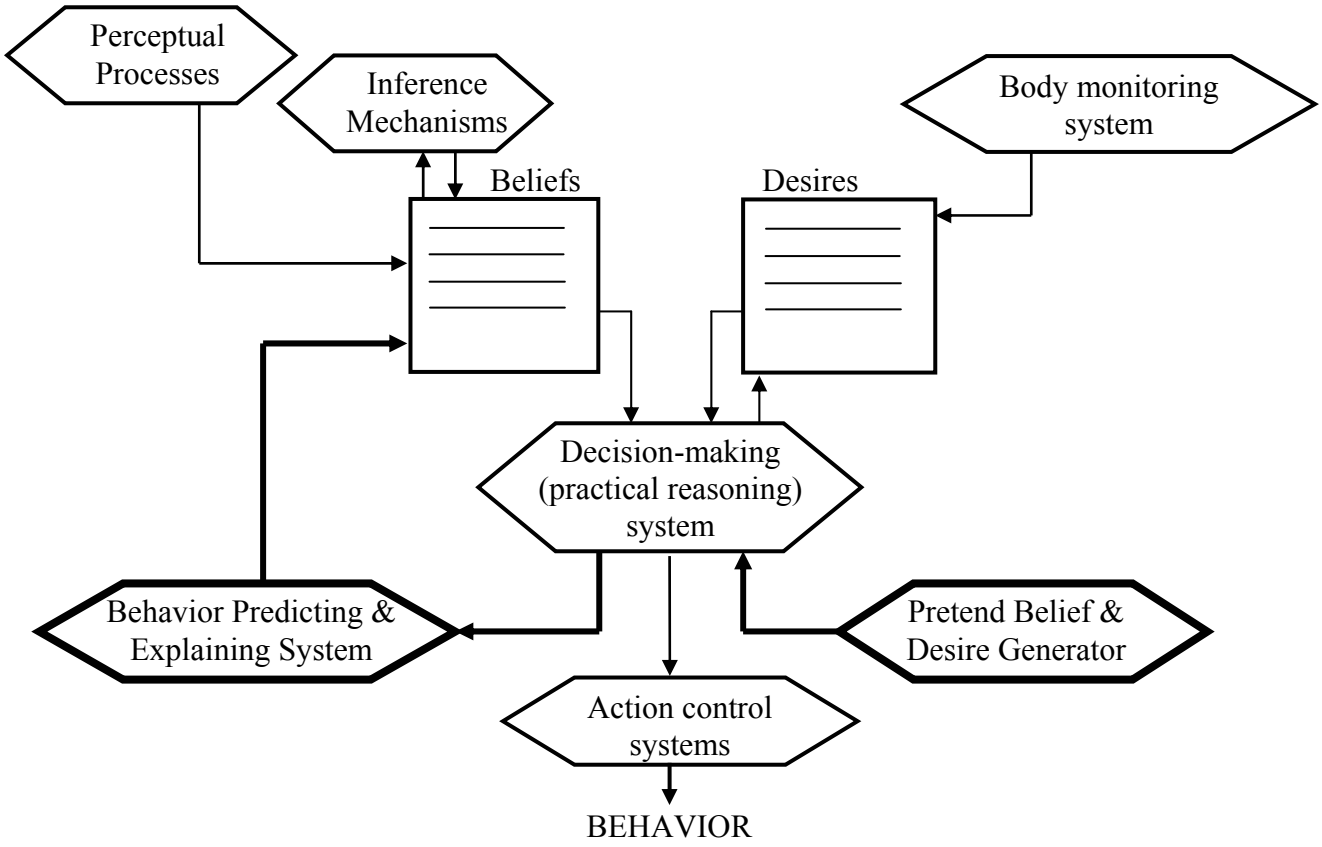

Figure 1

The process we just described takes the decision making system off-line and uses simulation to predict decisions. But much the same sort of process might be used to take the inference mechanism or other components of the mind off-line, and thus to make predictions about other sorts of mental processes. Some of the more enthusiastic defenders of simulation theory have suggested that *all* mindreading skills could be accomplished by something like this process of simulation, and thus that we need not suppose folk psychological theory plays *any* important role in mindreading. If this is right, then both functionalism and eliminativism are in trouble.[8]

---

[8] Robert Gordon is the most avid defender of view that all mindreading skills can be explained by simulation. Here is a characteristic passage:

> It is … uncanny that folk psychology hasn't changed very much over the millennia…. Churchland thinks this a sign that folk psychology is a bad theory; but it could be sign that it is no theory at all, not, at least, in the accepted sense of

4. Three accounts of mindreading: Information-rich, simulation based & hybrid

Simulation theorists and advocates of information-rich accounts of mindreading offer competing empirical theories about the mental processes underlying mindreading,[9] and much of the literature on the topic has been cast as a winner-takes-all debate between these two groups.[10] In recent years, however, there has been a growing awareness that mindreading is a complex, multifaceted phenomenon and that some aspects of mindreading might be subserved by information-poor simulation-like processes, while others are subserved by information-rich processes. This hybrid approach is one that we have advocated for a number of years (Stich & Nichols, 1995; Nichols, et al. 1996; Nichols & Stich, forthcoming), and in this section we'll give a brief sketch of the case in favor of the hybrid approach.[11] We'll begin by focusing on one important aspect of mindreading for which information-rich explanations are particularly implausible and a simulation-style account is very likely to be true. Well then take up two other aspects of mindreading where, we think, information-rich explanations are clearly to be preferred over simulation based explanations.

> (roughly) a system of laws implicitly defining a set of terms. Instead, it might be just the capacity for practical reasoning, supplemented by a special use of a childish and primitive capacity for pretend play. (1986, p. 71)

Of course, an eliminativist might object that the simulation theorist begs the question since the simulation account of decision prediction presupposes the existence of beliefs, desires and other posits of folk psychology, while eliminativists hold that these commonsense mental states not exist. Constructing a plausible reply to this objection is left as an exercise for the reader.

[9] Though Heal (1998) has argued that there is one interpretation of simulation theory on which it is true *a priori*. For a critique see Nichols & Stich (1998).

[10] Many of the important papers in this literature are collected in Davies & Stone (1995a & 1995b).

[11] We have also argued that some important aspects of mindreading are subserved by processes that can't be comfortably categorized as either information-rich or simulation-like. But since space is limited, we won't try to make a case for that here. See Nichols & Stich (forthcoming).

Inference Prediction: A Mindreading Skill Subserved by Simulation

One striking fact about the mindreading skills of normal adults is that we are remarkably good at predicting the inferences of targets, even their obviously *nondemonstrative* inferences. Suppose, for example, that Fred comes to believe that the President of the United States has resigned, after hearing a brief report on the radio. Who does Fred think will become President? We quickly generate the prediction that Fred thinks the Vice President will become President. We know perfectly well, and so, we presume, does Fred, that there are lots of ways in which his inference could be mistaken. The Vice President could be assassinated; the Vice President might resign before being sworn in as President; a scandal might lead to the removal of the Vice President; there might be a coup. It is easy to generate stories on which the Vice President would not become the new President. Yet we predict Fred's nondemonstrative inference without hesitation. And in most cases like this, our predictions are correct. Any adequate theory of mindreading needs to accommodate these facts.

Advocates of information-rich approaches to mindreading have been notably silent about inference prediction. Indeed, so far as we have been able to determine, no leading advocate of that approach has even *tried* to offer an explanation of the fact that we are strikingly good at predicting the inferences that other people make. And we're inclined to think that the reason for this omission is pretty clear. For a thorough going advocate of the information-rich approach, the only available explanation of our inference prediction skills is *more information*. If we are good at predicting how other people will reason, that must be because we have somehow acquired a remarkably good theory about how people reason. But that account seems rather profligate. To see why, consider the analogy between predicting inferences and predicting the grammatical intuitions of someone who speaks the same language that we do. To explain our success at this latter task, an advocate of the information-rich approach would have to say that we have a *theory* about the processes subserving grammatical intuition production in other people. But, as Harris (1992) pointed out, that seems rather far-fetched. A much simpler hypothesis is that we rely on *our own* mechanisms for generating linguistic intuitions, and having determined our own intuitions about a particular sentence, we attribute them to the target.

Harris's *argument from simplicity*, as we shall call it, played an important role in convincing us that a comprehensive theory of mindreading would have to invoke many different sorts of processes, and that simulation processes would be among them. However, we don't think that the argument from simplicity is the only reason to prefer a simulation-based account of inference prediction over an information-rich account.

Indeed, if the argument from simplicity were the only one available, a resolute defender of the information-rich approach might simply dig in her heels and note that the systems Mother Nature produces are often far from simple. There are lots of examples of redundancy and apparently unnecessary complexity in biological systems. So, the information-rich theorist might argue, the mere fact that a theory-based account of inference prediction would be less simple than a simulation style account is hardly a knock down argument against it. There is, however, another sort of argument that can be mounted against a information-rich approach to inference prediction. We think it is a particularly important argument since it can be generalized to a number of other mindreading skills, and thus it can serve as a valuable heuristic in helping us to decide which aspects of mindreading are plausibly treated as simulation based.

This second argument, which we'll call the *argument from accuracy*, begins with the observation that inference prediction is remarkably accurate over a wide range of cases, including cases that are quite different from anything that most mindreaders are likely to have encountered before. There is, for example, a rich literature in the "heuristics and biases" tradition in cognitive social psychology chronicling the ways in which people make what appear to be very bad inferences on a wide range of problems requiring deductive and inductive reasoning.[12] In all of this literature, however, there is

---

[12] Among the best known experiments of this kind are those illustrating the so-called *conjunction fallacy*. In one quite famous experiment, Kahneman and Tversky (1982) presented subjects with the following task.

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.
>
> Please rank the following statements by their probability, using 1 for the most probable and 8 for the least probable.
>
> (a) Linda is a teacher in elementary school.
> (b) Linda works in a bookstore and takes Yoga classes.
> (c) Linda is active in the feminist movement.
> (d) Linda is a psychiatric social worker.
> (e) Linda is a member of the League of Women Voters.
> (f) Linda is a bank teller.
> (g) Linda is an insurance sales person.
> (h) Linda is a bank teller and is active in the feminist movement.

no suggestion that people are bad at *predicting* other people's inferences, whether those inferences are good or bad. This contrasts sharply with the literature on desire attribution that we discuss below, where it is often remarked how surprising and *un*predictable people's desires and decisions are. Although it hasn't been studied systematically, we think it is quite likely that people typically predict others will make just those bad inferences that they would make themselves, even on problems that are quite different from any they have encountered before. If that is indeed the case, it poses a problem for information-rich accounts: How do ordinary mindreaders manage to end up with such an *accurate* theory about how people draw inferences – a theory which supports correct predictions even about quite unfamiliar sorts of inferences? The problem is made more acute by the fact that there are other sorts of mindreading tasks on which people do very badly. Why do people acquire the right theory about inference and the *wrong* theory about other mental processes? A simulation-based account of inference prediction, by contrast, has a ready explanation of our accuracy. On the simulation account, we are using the same inference mechanism for both *making* and *predicting* inferences, so it is to be expected that we would predict that other people make the same inferences we do.

Obviously, the argument from accuracy is a two edged sword. In those domains where we are particularly good at predicting or attributing mental states in unfamiliar cases, the argument suggests that the mindreading process is unlikely to be subserved by an information-rich process. But in those cases where we are *bad* at predicting or attributing mental states the argument suggests that the process is unlikely to be subserved by a *simulation* process. We recognize that there are various moves that might be made in response to the argument from accuracy, and thus we do not treat the argument as definitive. We do, however, think that the argument justifies a strong initial presumption that accurate mindreading processes are subserved by simulation-

---

In a group of naive subjects with no background in probability and statistics, 89% judged that statement (h) was more probable than statement (f) despite the obvious fact that one cannot be a *feminist* bank teller unless one is a *bank teller*. When the same question was presented to statistically sophisticated subjects – graduate students in the decision science program of the Stanford Business School – 85% gave the same answer! Results of this sort, in which subjects judge that a compound event or state of affairs is more probable than one of the components of the compound, have been found repeatedly since Kahneman and Tversky's pioneering studies, and they are remarkably robust. For useful reviews of research in the heuristics and biases tradition, see Kahneman et al. 1982, Nisbett & Ross 1980; Baron 2001; Samuels, Stich & Faucher 2001.

like processes and that inaccurate ones are not. And if this is right, then there is a strong presumption in favor of the hypothesis that inference prediction is simulation based.

Desire Attribution: A Mindreading Skill that Can't Be Explained by Simulation

Another quite central aspect of mindreading is the capacity to attribute desires to other people. Without that capacity we would not know what other people want, and we would be severely impaired in trying to predict or explain their actions. There are a number of processes that can give rise to beliefs about a target's desires. In some cases we use information about the target's verbal and non-verbal behavior (including their facial expressions) to determine what they want. In other cases we attribute desires on the basis of what other people say about the target. And in all likelihood a variety of other cues and sources of data are also used in the desire attribution process. It is our contention that these desire attribution skills do not depend on simulation but rather are subserved by information-rich processes. We have two quite different reasons for this claim.

First, desire attribution exhibits a pattern of systematic inaccuracy and that supports at least an initial presumption that the process is not simulation based. One very striking example comes from what is perhaps the most famous series of experiments in all of social psychology. Milgram (1963) had a "teacher" subject flip switches that were supposed to deliver shocks to another subject, the "learner" (who was actually an accomplice). For each mistake the learner made, the teacher was instructed to deliver progressively stronger shocks including one labeled "Danger: Severe Shock" and culminating in a switch labeled "450-volt, XXX". If the teacher subject expressed reservations to the experimental assistant, he was calmly told to continue the experiment. The result of the experiment was astonishing. A clear majority of the subjects administered *all* the shocks. People often find these results hard to believe. Indeed, the Milgram findings are so counterintuitive that in a verbal reenactment of the experiment, people still didn't predict the results (Bierbrauer 1973, discussed in Nisbett & Ross 1980, p. 121). One plausible interpretation of these findings is that in the Milgram experiment the instructions from the experimenter generated a desire to comply which, in most cases, overwhelmed the subject's desire not to harm the person they believed to be on the receiving end of the electric shock apparatus. The fact that people find the results surprising and that Bierbrauer's subjects did not predict them indicates an important limitation in our capacity to determine the desires of others.

There is a large literature in cognitive social psychology detailing many other

cases in which desires and preferences are affected in remarkable and unexpected ways by the circumstances subjects encounter and the environment in which they are embedded. The important point, for present purposes, is that people typically find these results surprising and occasionally quite unsettling, and the fact that they are surprised (even after seeing or getting a detailed description of the experimental situation) indicates that the mental mechanisms they are using to predict the subjects' desires and preferences are systematically inaccurate. Though this is not the place for an extended survey of the many examples in the literature, we can't resist mentioning one of our favorites.[13]

In a recent study, Loewenstein and Adler (1995) looked at the ability of subjects to predict *their own* preferences when those preferences are influenced by a surprising and little known effect. The effect that Loewenstein & Adler exploit is the *endowment effect*, a robust and rapidly appearing tendency for people to set a significantly higher value for an object if they actually own it than they would if they did not own it (Thaler, 1980). Here is how Loewenstein & Adler describe the phenomenon.

> In the typical demonstration of the endowment effect … one group of subjects (sellers) are endowed with an object and are given the option of trading it for various amounts of cash; another group (choosers) are not given the object but are given a series of choices between getting the object or getting various amounts of cash. Although the objective wealth position of the two groups is identical, as are the choices they face, endowed subjects hold out for significantly more money than those who are not endowed (Loewenstein and Adler, 1995, pp. 929-930).

In an experiment designed to test whether "unendowed" subjects could predict the value they would set if they were actually to own the object in question, the experimenter first allowed subjects (who were members of a university class) to examine a mug engraved with the school logo. A form was then distributed to approximately half of the subjects, chosen at random, on which they were asked "to imagine that they possessed the mug on display and to predict whether they would be willing to exchange the mug for various amounts of money" (Loewenstein and Adler 1995, p. 931). When the subjects who received the form were finished filling it out, *all* the subjects were presented with a mug and given a second form with instructions analogous to those on the prediction form. But on the second form it was made clear that they actually could exchange the mug for cash, and that the choices they made on this second form would determine how much money they might get. "Subjects were

[13] For an excellent review of the literature, see Ross & Nisbett (1991).

told that they would receive the option that they had circled on one of the lines – which line had been determined in advance by the experimenter" (Loewenstein and Adler, 1995, p. 931). The results showed that subjects who had completed the first form substantially underpredicted the amount of money for which they would be willing to exchange the mug. In one group of subjects, the mean predicted exchange price was $3.73, while the mean actual exchange price for subjects (the *same* subjects who made the prediction) was $5.40! Moreover, there seemed to be an "anchoring effect" in this experiment which depressed the actual exchange price, since the mean actual exchange price for subjects who did not make a prediction about their own selling price was even higher at $6.46. Here again we find that people are systematically inaccurate at predicting the effect of the situation on desires, and in this case the desires they fail to predict are their own! If these desire predictions were subserved by a simulation process, it would be something of a mystery why the predictions are systematically inaccurate. But if, as we believe, they are subserved by an information-rich process, the inaccuracy can be readily explained. The theory or body of information that guides the prediction simply does not have accurate information about the rather surprising mental processes that give rise to these desires.

Our second reason for thinking that the mental mechanisms subserving desire attribution use information-rich processes rather than simulation is that it is hard to see how the work done by these mechanisms *could* be accomplished by simulation. Indeed, so far as we know, simulation theorists have made only one proposal about how some of these desire detection tasks might be carried out, and it is singularly implausible. The proposal, endorsed by both Gordon (1986) and Goldman (1989) begins with the fact that simulation processes like the one sketched in Figure 1 can be used to make behavior predictions, and goes on to suggest that they might also be used to generate beliefs about the desires and beliefs that give rise to observed behavior by exploiting something akin to the strategy of analysis-by-synthesis (originally developed by Halle & Stevens (1962) for phoneme recognition). In using the process in Figure 1 to predict behavior, hypothetical or "pretend" beliefs and desires are fed into the mindreader's decision making system (being used "off-line" of course), and the mindreader predicts that the target would do what the mindreader would decide to do, given those beliefs and desires. In an analysis-by-synthesis account of the generation of beliefs about desires and beliefs, the process is, in effect, run backwards. It starts with a behavioral episode that has already occurred and proceeds by trying to find hypothetical beliefs and desires which, when fed into the mindreader's decision mechanism, will produce a decision to perform the behavior we want to explain.

An obvious problem with this strategy is that it will generate too many candidates, since typically there are endlessly many possible sets of beliefs and desires

that might lead the mindreader to decide to perform the behavior in question. Gordon is well aware of the problem, and he seems to think he has a solution:

> No matter how long I go on testing hypotheses, I will not have tried out all candidate explanations of the [target's] behavior. Perhaps some of the unexamined candidates would have done at least as well as the one I settle for, if I settle: perhaps indefinitely many of them would have. But these would be 'far fetched', I say intuitively. Therein I exhibit my inertial bias. The less 'fetching' (or 'stretching', as actors say) I have to do to track the other's behavior, the better. I tend to feign only when necessary, only when something in the other's behavior doesn't fit…. This inertial bias may be thought of as a 'least effort' principle: the 'principle of least pretending'. It explains why, other things being equal, I will prefer the less radical departure from the 'real' world - i.e. from what I myself take to be the world. (Gordon 1986, p. 164).

Unfortunately, it is not at all clear what Gordon has in mind by an inertial bias against "fetching". The most obvious interpretation is that attributions are more "far-fetched" the further they are, on some intuitive scale, from one's own mental states. But if that's what Gordon intends, it seems clear that the suggestion won't work. For in many cases we explain behavior by appealing to desires or beliefs (or both) that are *very* far from our own. I might, for example, explain the cat chasing the mouse by appealing to the cat's desire to eat the mouse. But there are indefinitely many desires that would lead me to chase a mouse that are intuitively much closer to my actual desires than the desire to eat a mouse! Simulation theorists have offered no other proposal for narrowing down the endless set of candidate beliefs and desires that the analysis-by-synthesis strategy would generate, and without some plausible solution to this problem the strategy looks quite hopeless. So it is not surprising that accounts of this sort have largely disappeared from the simulation theory literature over the last decade. And that, perhaps, reflects at least a tacit acknowledgement, on the part of simulation theorists, that desire attribution can only be explained by appealing to information-rich processes.

Discrepant Belief Attribution: Another Mindreading Skill that Can't Be Explained by Simulation

Yet another important aspect of mindreading is the capacity to attribute beliefs that we ourselves do not hold – *discrepant beliefs* as they are sometimes called. There are a number of processes subserving discrepant belief attribution, some relying on beliefs about the target's perceptual states, others exploiting information about the target's verbal behavior, and still others relying on information about the target's non-verbal

behavior. All of these, we suspect, are subserved by information-rich mechanisms, rather than by a mechanism that uses simulation. Our reasons are largely parallel to the ones we offered for desire attribution. First, there is abundant evidence that the discrepant belief attribution system exhibits systematic inaccuracies of the sort we would expect from an information-rich system that is not quite rich enough and does not contain information about the process generating certain categories of discrepant beliefs. Second, there is no plausible way in which prototypical simulation mechanisms could do what the discrepant belief attribution system does.

One disquieting example of a systematic failure in discrepant belief attribution comes from the study of belief perseverance. In the psychology laboratory, and in everyday life, it sometimes happens that people are presented with fairly persuasive evidence (e.g. test results) indicating that they have some hitherto unexpected trait. In light of that evidence people typically form the belief that they do have the trait. What will happen to that belief if, shortly after this, people are presented with a convincing case discrediting the first body of evidence? Suppose, for example, they are convinced that the test results they relied on were actually someone else's, or that no real test was conducted at all. Most people expect that the undermined belief will simply be discarded. And that view was shared by a generation of social psychologists who duped subjects into believing all sorts of things about themselves, often by administering rigged psychological tests, observed their reactions, and then "debriefed" the subjects by explaining the ruse. The assumption was that no enduring harm could be done because once the ruse was explained the induced belief would be discarded. But in a widely discussed series of experiments, Ross and his co-workers have demonstrated that this is simply not the case. Once a subject has been convinced that she has a trait, showing her that the evidence that convinced her was completely phony does not succeed in eliminating the belief. (Nisbett & Ross 1980, p. 175-179) If the trait in question is being inclined to suicide, or being "latently homosexual," belief perseverance can lead to serious problems. The part of the discrepant belief attribution system that led both psychologists and everyone else to expect that these discrepant beliefs would be discarded after debriefing apparently has inaccurate information about the process of belief perseverance and thus it leads to systematically mistaken belief attributions.

Another example, with important implications for public policy, is provided by the work of Loftus (1979) and others on the effect of "postevent interventions" on what people believe about events they have witnessed. In one experiment subjects were shown a film of an auto accident. A short time later they were asked a series of questions about the accident. For some subjects, one of the questions was, "How fast was the white sports car traveling when it passed the barn while traveling along the

country road?" Other subjects were asked, "How fast was the white sports car traveling while traveling along the country road?" One week later all the subjects were asked whether they had seen a barn. Though there was no barn in the film that the subjects had seen, subjects who were asked the question that mentioned the barn were five times more likely to believe that they had seen one. In another experiment, conducted in train stations and other naturalistic settings, Loftus and her students staged a "robbery" in which a male confederate pulled an object from a bag that two female students had temporarily left unattended and stuffed it under his coat. A moment later, one of the women noticed that her bag had been tampered with and shouted, "Oh my God, my tape recorder is missing." She went on to lament that her boss had loaned it to her and that it was very expensive. Bystanders, most of whom were quite cooperative, were asked for their phone numbers in case an account of the incident was needed for insurance purposes. A week later, an "insurance agent" called the eyewitnesses and asked about details of the theft. Among the questions asked was "Did you see the tape recorder?" More than half of the eyewitnesses remembered having seen it, and nearly all of these could describe it detail – this despite the fact that *there was no tape recorder*! On the basis of this and other experiments, Loftus concludes that even casual mention of objects that were not present or of events that did not take place (for example, in the course of police questioning) can significantly increase the likelihood that the objects or events will be incorporated into people's beliefs about what they observed. A central theme in Loftus's work is that the legal system should be much more cautious about relying on eyewitness testimony. And a major reason why the legal system is *not* as cautious as it should be is that our information-driven discrepant belief attribution system lacks information about the postevent processes of belief formation that Loftus has demonstrated.

As in the case of desire attribution, we see no plausible way in which the work done by the mental mechanisms subserving discrepant belief attribution *could* be accomplished by simulation. Here again, the only proposal that simulation theorists have offered is the analysis-by-synthesis account, and that strategy won't work any better for belief attribution than it does for desire attribution.

5. Conclusion

In the previous section we sketched some of the reasons for accepting a hybrid account of mindreading in which some aspects of that skill are explained by appeal to information-rich processes while other aspects are explained by simulation. Though we only looked at a handful of mindreading skills, we have argued elsewhere (Nichols & Stich, forthcoming) that much the same pattern can be found more generally.

Mindreading is a complex and multifaceted phenomenon, many facets of which are best explained by an information rich approach, while many other facets are best explained by simulation. If this is correct, it presents both functionalists and eliminativists with some rather awkward choices. Functionalists, as we have seen, hold that the meaning of ordinary mental state terms is determined by folk psychology, and eliminativists typically agree. In Section 2 we argued that functionalism is most plausible if folk psychology is taken to be the information-rich theory that subserves mindreading. But now it appears that only *parts* of mindreading rely on an information-rich theory. Should functionalists insist that the theory underlying these aspects of mindreading fixes the meaning of mental state terms, or should they retreat to the platitude account of folk psychology? We are inclined to think that whichever option functionalists adopt, their theory will be less attractive than it was before it became clear that the platitude approach and the mindreading approach would diverge, and that only part of mindreading relies on folk psychology.

References

Armstrong, D. (1968). *A Materialist Theory of the Mind*. New York: Humanities Press.

Baron, J. (2001). *Thinking and Deciding*. Third edition. Cambridge: Cambridge University Press.

Bierbrauer, G. 1973. *Effect of Set, Perspective, and Temporal Factors in Attribution*, unpublished doctoral dissertation, Stanford University, 1973.

Block, N. (1994). Functionalism. In S. Guttenplan (ed.), *A Companion to the Philosophy of Mind*. Oxford: Blackwell. 323-332.

Churchland, P. (1981). Eliminative materialism and propositional attitudes. *Journal of Philosophy*, 78, 67-90. Reprinted in W. Lycan (ed.), *Mind and Cognition*. Oxford: Blackwells, 1990. 206-223. Page reference is to the Lycan volume.

Davies, M. & Stone, T. (1995a). *Folk Psychology*. Oxford: Blackwell.

Davies, M. & Stone, T. (1995b). *Mental Simulation*. Oxford: Blackwell.

Devitt, M. (1996). *Coming to Our Senses: A Naturalistic Program for Semantic Localism*. Cambridge: Cambridge University Press.

Ekman, P. 1985. *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. New York: W. W. Norton & Co.

Fodor, J. & LePore, E. (1992). *Holism: A Shopper's Guide*. Oxford: Blackwell.

Fodor, J. & Chihara (1965). Operationalism and ordinary language. *American Philosophical Quarterly*, 2, 4. Reprinted in J. Fodor, *Representations*. Cambridge, MA: MIT Press, 1981. 35-62.

Gigerenzer, G., Todd, P. & the ABC Research Group (1999). *Simple Heuristics that Make Us Smart*. Oxford: Oxford University Press.

Goldman, A. (1989). Interpretation psychologized. *Mind and Language* 4, 161-185.

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

Gopnik, A. & Wellman, H. (1994). The Theory-Theory. In L. Hirschfeld & S. Gelman (eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture*. New York: Cambridge University Press. 257-293.

Gordon, R. (1986). Folk psychology as simulation. *Mind and Language*, 1, 158-170. Reprinted in Davies & Stone (1995a). Page reference is to Davies & Stone.

Halle, M. & Stevens, K. (1962). Speech recognition: A Model and a program for research. In Fodor & Katz (eds.) *The structure of language: Readings in the philosophy of language*. Englewood Cliffs, NJ: Prentice-Hall.

Harris, P. (1991). The work of the imagination. In A. Whiten (ed.), *Natural theories of mind*. Oxford: Blackwell.

Hayes, P. (1985). The second naive physics manifesto. In J. Hobbs & R. Moore (eds), *Formal Theories of the Commonsense World*. Norwood, NJ: Ablex. 1-36.

Heal, J. (1986). Replication and functionalism. In J. Butterfield (ed.), *Language, Mind and Logic*. Cambridge: Cambridge University Press. 135-150.

Heal, J. (1998). Co-cognition and Off-line simulation: Two ways of understanding the simulation approach. *Mind & Language*, 13, 477-498.

Hempel, C. (1964). The theoretician's dilemma: A study in the logic of theory construction. In C. Hempel, *Aspects of Scientific Explanation*. New York: The Free Press. 173-226.

Kahneman, D., Slovic, P. and Tversky, A. (eds.), (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

Kahneman, D. and Tversky, A. (1982). The psychology of preferences. *Scientific American*, vol. 246 (1), 160-173.

Lewis, D. (1970). How to define theoretical terms. *Journal of Philosophy*, 67,17-25.

Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy*, 50, 249-58. Reprinted in N. Block (ed.), *Readings in the Philosophy of Psychology*, Vol. I. Cambridge, MA: Harvard University Press. 207-215. Page references are to the Block volume.

Loewenstein, G. & Adler, D. (1995). A bias in the prediction of tastes. *The Economic Journal: The Quarterly Journal of the Royal Economic Society*, 105, 929-937.

Loftus, E. (1979). *Eyewitness Testimony*. Cambridge, MA: Harvard University Press.

Lycan, W. (1994). Functionalism. In S. Guttenplan (ed.), *A Companion to the Philosophy of Mind*. Oxford: Blackwell. 317-323.

McCloskey, M. (1983). Intuitive physics, *Scientific American*, 248, 4, 122-129.

Milgram, S. (1963). Behavioral study of obedience, *Journal of Abnormal and Social Psychology*, 67, 371-78.

Nichols, S. & Stich, S. (1998). Rethinking co-cognition: A reply to Heal. *Mind & Language*, 13, 499-512.

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

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

Nisbett, R. & Ross, L. (1980). *Human Inference*. Englewood Cliffs, NJ: Prentice-Hall.

Putnam, H. (1960). Minds and machines. In S. Hook (ed.), *Dimensions of Mind*. New York: New York University Press. Pp. 138-164.

Ross, L. & Nisbett, R. (1991). *The Person and the Situation: Perspectives of Social Psychology.*

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

Samuels, R., Stich, S. & Faucher, L. (2001). Reasoning and rationality. In I. Niiniluoto, M. Sintonen, & J. Wolenski (eds.), *Handbook of Epistemology*. Dordrecht: Kluwer. Pp. 1-50.

Scholl, B. & Leslie, A. (1999). Modularity, development, and 'Theory of Mind'. *Mind & Language*, 14, 131-153.

Sellars, W. (1956). Empiricism and the philosophy of mind. In H. Feigl & M. Scriven (eds.), *The Foundations of Science and the Concepts of Psychology and Psychoanalysis: Minnesota Studies in the Philosophy of Science*, Vol. 1. Minneapolis: University of Minnesota Press. 253-329.

Stich, S. (1996). *Deconstructing the mind*. Oxford: Oxford University Press.

Stich, S. & Nichols, S. (1995). Second thoughts on simulation. In Davies & Stone (1995,b). 86-108.

Stich, S. & Ravenscroft, I. (1994). What *is* folk psychology? *Cognition*, 50, 447-468. Reprinted in Stich (1996).

Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior and Organization*, 1, 39-60.

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