This paper was published in In *Mental Simulation: Evaluations and Applications*, eds. M. Davies and T. Stone. Oxford: Basil Blackwell, 1995, 87-108.

Archived at Website for the Rutgers University Research Group on Evolution and Higher Cognition.

Second Thoughts on Simulation

Stephen Stich
Department of Philosophy and Center for Cognitive Science
Rutgers University
New Brunswick, NJ 08901
stich@ruccs.rutgers.edu

and

Shaun Nichols
Department of Philosophy
College of Charleston
Charleston, SC 29424
nichols@cofc.edu

Contents

The essays in this volume make it abundantly clear that there is no shortage of disagreement about the plausibility of the simulation theory. As we see it, there are at least three factors contributing to this disagreement. In some instances the issues in dispute are broadly empirical. Different people have different views on which theory is favored by experiments reported in the literature, and different hunches about how future experiments are likely to turn out. In 3.1 and 3.3 we will consider two cases that fall under this heading. With a bit of luck these disputes will be resolved as more experiments are done and more data become available. Faulty arguments are a second source of disagreement. In 3.2 and 3.4 we will set out two dubious arguments advanced by our critics and try to explain exactly why we think they are mistaken. The third source of disagreement is terminological. Terms like "theory-theory," "simulation theory" and a number of others are often not clearly defined, and they are used in different ways by different authors. (Worse yet, we suspect they are sometimes used in different ways by a single author on different occasions). Thus it is sometimes the case that what appears to be a substantive disagreement turns out to be simply a verbal dispute. Moreover, since the labels "theory-theory" and "simulation theory" are each used to characterize a broad range of theories, it may well turn out that some of the theories falling under both headings are correct. In Sections 1 and 2, we will set out a variety of different views for which the labels "theory-theory" and "simulation theory" might be used. As we proceed we'll point out a number of places where disagreements diminish when distinctions among different versions of the theory-theory and the simulation theory are kept clearly in mind.

1. What is the Theory-Theory?

The central idea shared by all versions of the "theory-theory" is that the processes underlying the production of most

predictions, intentional descriptions, and intentional explanations of people's behavior exploit an internally represented body of information (or perhaps mis-information) about psychological processes and the ways in which they give rise to behavior. We call this body of information "folk psychology". If the theory-theory is correct, then people's predictions and explanations of human behavior are subserved by a mechanism quite similar to the one that is widely supposed to subserve commonsense predictions and explanations about the behavior of ordinary physical objects. Those predictions, too, invoke an internally represented body of information -- in this case a "folk physics." In Figure 1, we've sketched what we've found to be a useful "boxological" rendition of the theory-theory.[(1)]

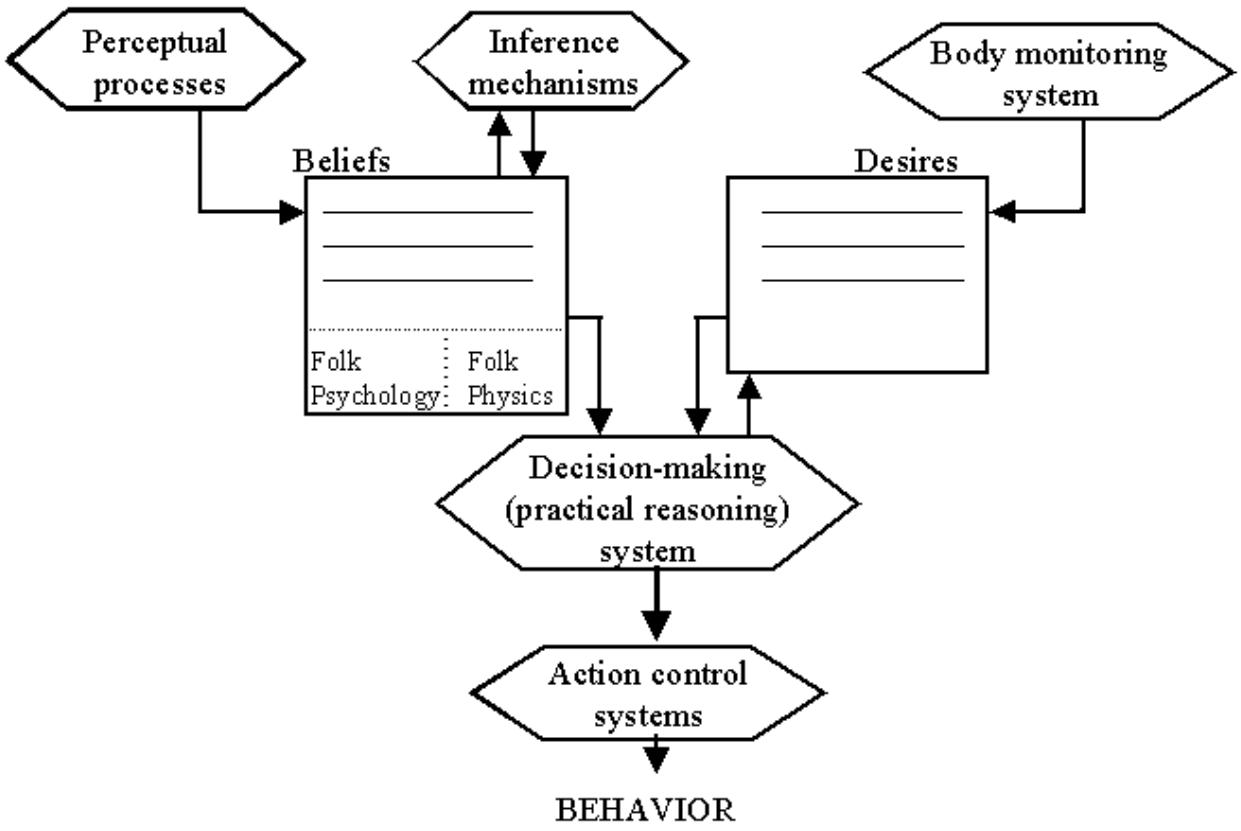

Figure 1

The basic tenets of the theory-theory can be developed in a variety of ways. In our earlier paper[(2)] we focused on differing accounts of the way in which the information in the "folk psychology box" might be stored. One possibility is that the information in the box is stored in a sentence-like or rule based system. But it is also possible that some quite different strategy of storage is exploited -- a connectionist network, perhaps, or something along the lines of the "mental models" championed by Johnson-Laird.[(3)] There are, however, lots of other distinctions that might be noted, most of which are compatible with a variety of views on how the information is stored. Here is a brief catalogue:

i) A number of authors expect that folk psychology will include a significant number of lawlike or nomological generalizations. Others expect that folk psychology will contain little or nothing that resembles the laws of mature sciences. Rather, they suppose, it will consist of rules of thumb, prototypical cases, or other sorts of claims that aren't even prima facie lawlike.[(4)]

ii) Some authors believe that folk psychology invokes "theoretical" entities or processes which are quite distinct from the more readily observable events and processes that folk psychology will be called on to predict and explain. They also suppose that the theoretical constructs of folk psychology will be "closely or `lawfully' interrelated with one another."[(5)]

If this is correct, then folk psychology will bear a strong resemblance to the standard philosophical portrait of scientific theories in domains like physics and chemistry. But, of course, there are lots of domains of commonsense knowledge in which it is rather implausible to suppose that the mentally represented "knowledge structure" includes theoretical constructs linked together in lawlike ways. Knowledge of cooking, or of current affairs are likely candidates here, as is the knowledge that underlies our judgements about what is polite and impolite in our culture. And it is entirely possible that folk psychological knowledge will turn out to resemble the knowledge structures underlying cooking or politeness judgements rather than the knowledge structures that underlie the scientific predictions and explanations produced by a competent physicist or chemist. If this is how things turn out, then as we use the term "theory-theory" it would still be the case that the theory-theory is true. Other authors prefer to use the term "theory-theory" more narrowly. On their usage, an account of the knowledge structures underlying ordinary predictions and explanations of behavior which does not include theoretical constructs and lawlike generalizations would not count as a version of the "theory-theory."

Now, so far as we can see, there is no substantive issue at stake in the choice of terminology. Those who adopt our very inclusive use of the term "theory-theory" will need to introduce some more specialized labels to distinguish those versions of the theory-theory on which folk psychology includes laws and theoretical constructs from those on which it does not. Those who prefer a more restrictive use of "theory-theory" will have to introduce some new label for theories which hold that the stuff in the "folk psychology box" does not include lawlike generalizations. Along with many other participants in the debate, we are inclined to think that the theory-theory (construed broadly) and the off-line simulation theory (as we characterized it in FPSTT) are the only two plausible proposals that have been made about the mechanisms underlying people's ability to predict other people's behavior. They are, as they say, the only games in town. Moreover, as we characterize them, the two are mutually exclusive. Thus an argument against the theory-theory (construed broadly) would lend considerable support to the off-line simulation theory. But, and this is the crucial point here, an argument against (what we would describe as) some particular version of the theory-theory will not lend much support to the off-line simulation theory. The theory-theory (construed broadly) does not insist that folk psychology includes lawlike generalizations. So even if it could be shown that people do not exploit lawlike generalizations in predicting and explaining other people's behavior, this would not show that the theory-theory is wrong, and it would not provide any significant degree of support for the simulation theory.

iii) Yet another issue on which theory-theorists may differ is the extent to which the contents of the "folk-psychology box" are accessible to consciousness. According to Goldman, "[i]n the philosophical literature it has been widely assumed that it should be easy to formulate the principles of folk psychology because they are platitudes..."[(6)] Unfortunately, he provides no references to support this contention, and we are not at all sure that he's correct. It is indeed the case that in a widely discussed paper David Lewis claims that the meaning of mental state terms can be captured by attending to the commonsense psychological "platitudes" that are obvious to everyone.[(7)] But it is far from clear that Lewis or those who followed him thought that these very same platitudes were what people used to predict and explain each other's behavior. Indeed, on our reading of Lewis's paper, he offers no view at all about this. Thus we are rather dubious of Goldman's claim that the theory-theorist who maintains that the principles of folk psychology are "tacit, no more accessible than internalized rules of grammar"[(8)] is "shift[ing] away from the original form of the theory-theory." But of course this is just a scholarly quibble. The important point is that if it turns out that the contents of the "folk psychology box" are no more accessible than the rules of grammar, this will not show that the theory-theory (broadly construed) is mistaken, nor will it lend any support to the off-line simulation theory. And on this point, it seems, Goldman agrees with us.

iv) The final distinction we'll consider turns on how folk psychology is acquired. According to Goldman, the "standard" assumption has been that "folk psychological platitudes are culturally produced and culturally transmitted." In support of this claim, he quotes Churchland who maintains that "All of us learn [the folk psychological] framework (at mother's knee, as we learn our language)...."[(9)] But if language learning is the appropriate analogy, then it is very odd indeed to protest that "few children have mothers that utter [folk psychological] platitudes." For exactly the same could be said about the principles of grammar that children learn when they learn their language. Few children have mothers who utter grammatical rules. Indeed for the last twenty five years or so the dominant view in generative linguistics has been that language learning is subserved by a special purpose language acquisition mechanism, and that there are strong innate constraints on the sort of grammars that this mechanism can produce. So if acquiring folk psychology was standardly presumed to be similar to language acquisition, then the "standard" assumption must have been that there is a special purpose folk-psychology acquisition device, and that there are strong innate constraints on the sort of folk psychological theory that the acquisition device can produce. Once again, however, this is little more than a scholarly quibble. The important point is the one that Goldman grants. The theory-theory is not committed to any particular account of how folk psychology is acquired, and thus "the theory-theory cannot be ruled out by the mystery of

acquisition." Nor, we would add, can the off-line simulation theory be supported.

2. What is the Simulation Theory?

The basic idea of what we call the "off-line simulation theory" is that in predicting and explaining people's behavior we take our own decision making system "off-line," supply it with "pretend" inputs that have the same content as the beliefs and desires of the person whose behavior we're concerned with, and let it make a decision on what to do. Figure 2 is a boxological rendition of the essential points of the off-line simulation theory.[(10)]

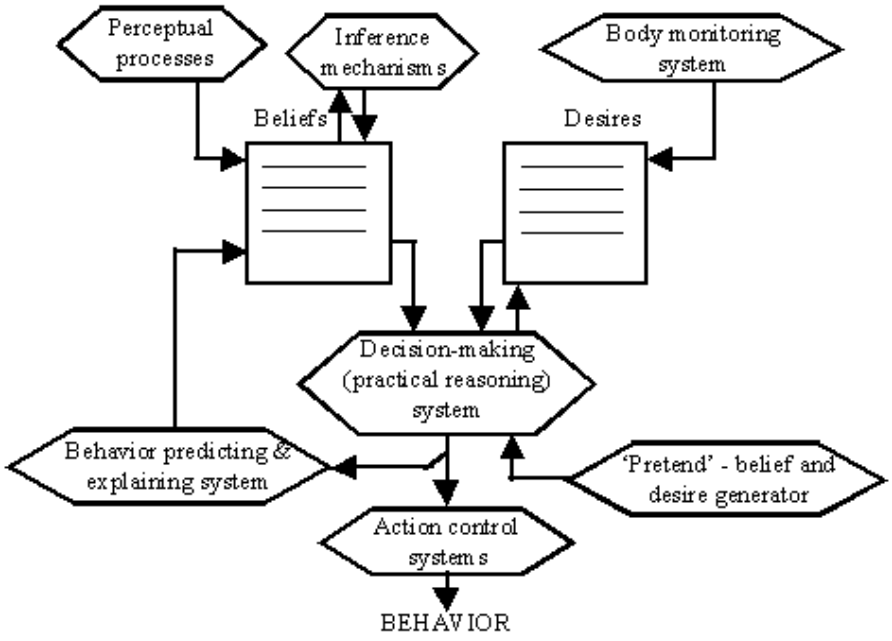

Figure 2

Some of the contributors to this volume share this conception of the simulation theory. But there are also some notable exceptions.

2.1. Gordon's Version of the Simulation Theory

On the terminology that Gordon prefers, the "off-line" picture sketched in Figure 2, is only an "ancillary hypothesis" about the mechanism underlying simulation, albeit a "very plausible" one. So what we call the "off-line simulation theory" is, for him, only one version of the simulation theory. What is essential for all versions of the simulation theory, Gordon tells us, is that we explain and predict behavior by "imaginative identification" - "that is, [we] use our imagination to identify with others in order to explain or predict their behavior"[(11)] Unfortunately, we are not at all sure we understand what Gordon means in these quotes and in other, similar, passages. We can think of at least three interpretations. The first is that all cases of explanation and prediction of behavior are accompanied by conscious imagery -- we imagine ourselves in the other person's situation in much the same way that we imagine ourselves walking through our house when we are asked to say how many windows are in our house.[(12)] A second interpretation is identical with the first, except that "all cases" is replaced by "some cases." We rather doubt that either of these is the

right interpretation, however. For on the first interpretation, what Gordon calls the "simulation theory" would be patently false. It is conceded on all sides that there are many cases in which the prediction and explanation of people's behavior is not accompanied by imagery or any other salient phenomenology. On the second interpretation, by contrast, the "simulation theory" is patently true, though it is not very exciting, and it is not incompatible with the theory-theory. A third interpretation, and the one which we take to be the most likely, is that Gordon uses locutions like "identification in imagination" and "imaginative identification with the other" as labels for a special sort of mental act or process which may be (but need not be) accompanied by conscious imagery. But if we are right, if this is how he uses these locutions, then he owes us a much more detailed account of this special sort of mental act or process. Without such an account, locutions like "imaginative identification with the other" are just unexplained jargon. To put the matter bluntly, if the third interpretation is the correct one, then, pending further clarification, we are not at all sure we understand what Gordon is talking about.

2.2. Harris's Version of the Simulation-Theory

In Harris's paper,[(13)] there is an important proposal of a way to understand "simulation" which is significantly different from the "off-line simulation" theory we set out in FPSTT. Here is how Harris spells out the idea:

> [C]onsider the following thought-experiment. I recruit your help, explaining that I am carrying out a two-part psycholinguistic study. I have presented English speakers with grammatical and ungrammatical sentences and they have made their judgements about which is which. I ask you to predict what decision most people came to about each sentence. Your hit rate turns out to be very high. In almost every case you can tell me whether the majority judged the sentence to be grammatical or ungrammatical. Moreover, when I ask you to explain your predictions you do so by indicating deviant constructions or morphemes in the ungrammatical sentences, something that speakers in the first part of the study also did. How are your predictions so accurate? The most plausible answer is that you read each sentence, asked yourself whether it sounded grammatical or not, and assumed that other English speakers would make the same judgments for the same reasons. The proposal that you have two distinct tacit representations of English grammar, a first-order representation that you deploy when making your own judgments, and a metarepresentation (i.e. a representation of other people's representations) that you deploy in predicting the judgements made by others, so designed as to yield equivalent judgements, strains both credulity and parsimony.[(14)]

We have no quarrel at all with Harris's account of how we would go about predicting another speaker's grammatical judgments. We agree that it would be singularly implausible to suppose that English speakers have two distinct tacit representations of English grammar. Moreover, we think Harris is quite right when he goes on to suggest that the sort of predictive strategy he is sketching generalizes to lots of other cases. Suppose, for example, that we were asked to predict what one of our Rutgers colleagues would say when asked: "Who is the President of Rutgers University?" Or suppose we were asked to predict what our wives would say when asked: "Who was the third President of the United States?" In both cases, we suspect, we would proceed by first answering the question for ourselves -- recalling who we think is the President of Rutgers or who we think was the third President of the U.S. Then, since we assume that our colleagues (in the first case) and our wives (in the second case) believe the same things we do on questions like this, we would predict that they would say the same thing we would.

In all of these cases, it would be perfectly natural to say that a sort of "simulation" is being exploited. But it is important to see that this is not the sort of "off-line" simulation sketched in Figure 2. Indeed, the processes that take place in these cases occur prior to the point at which the off-line simulation process might be supposed to kick in. To see the point, let's consider these cases a bit more closely. In none of them is the "simulation" actually being used to predict behavior. Rather, it is being used to figure out what the targets of the various predictions believe. In the Presidents cases, we figure out what the targets believe by determining what we ourselves believe, and then attributing the same belief to them. In the grammar case, we put ourselves in a situation similar to the one the target person is in -- we listen to the sentence, just as the target did -- and note what we come to believe. Then, once again, we attribute the same belief to the target. Having determined what the target believes, we can then proceed to predict what he or she will do. And it is here that either the processes posited by the theory-theory or the processes posited by the off-line simulation theory might be thought to operate. For these two theories provide different accounts of how we go from information about the beliefs, desires and other mental states of the target person to predictions about the target's behavior. On the off-line simulation account, the information about the target's belief (along, perhaps, with some other information about the target's desires and further beliefs) is fed into our own decision making system. That system makes a decision which, rather than being acted on is transformed into a prediction and reported to the "belief box". On the theory-theory, the

principles of our mentally represented folk psychological theory along with information about the beliefs and desires of the target are used to infer a prediction about how the target person will behave.

So our position on Harris's proposal is as follows. First, we agree that he has given an extremely plausible account of the way in which we sometimes figure out what someone else believes. Second, we grant that this process, in which we first determine what we ourselves believe, and then attribute the same belief to someone else, might perfectly well be described as a kind of "simulation". (As we find ourselves saying over and over again, there is no point in arguing who gets to keep the word.) However, this sort of simulation is quite different from the sort of off-line simulation described in FPSTT. It is not a process which results in predictions of behavior, and it is compatible with both the theory-theory and the off-line simulation theory, each of which provides an account of how we use information about a person's beliefs, desires and other mental states in producing predictions about that person's behavior. We propose to call the process Harris has described "type-1 Harris simulation." We have chosen this rather awkward label because we think that there is another, more controversial sort of simulation, which Harris may also have in mind, and which needs to be distinguished from type-1 Harris simulation and also from off-line simulation.

To explain this third sort of simulation, let us begin with a Harris-style thought experiment. Suppose we tell you the following story:

> Sven believes that all Italians like pasta. Sven is introduced to Maria, and he is told that she is Italian.

Now we ask you to make a prediction: If Sven is asked: "Does Maria like pasta?" what will Sven say? How do you go about making this prediction? One hypothesis, a natural extension of the theory-theory, is that your folk psychology includes a component specifying how people form beliefs from other beliefs - one might think of this component as a tacit theory of reasoning. From this theory, along with the premises about Sven, you infer that Sven will come to believe that Maria likes pasta. The process is analogous to the one in which you use principles of folk physics and premises about the current location of certain physical objects and the forces acting upon them to infer a conclusion about the future location of the objects. Another hypothesis, more in the spirit of simulation theories, is that you have the capacity to feed pretend or hypothetical inputs into your own inference mechanism, and then allow it to churn away as it normally does and draw appropriate conclusions. In the case at hand, the pretend inputs would be:

> All Italians like pasta.

and

> Maria is an Italian.

And the conclusion would be:

> Maria likes pasta.

This conclusion is not simply fed into your belief-box, however. For if it were, you would end up believing that Maria likes pasta (even if, unlike Sven, you believe that many Italians are not at all keen on pasta.) Rather, the conclusion churned out by your inference mechanism is attributed to Sven. You come to believe that Sven will come to believe that Maria likes pasta. So there must be some mechanism which takes the output of your inference-box and embeds it in a belief-sentence before it is fed into your belief-box. We propose to call the inference simulation process just described "type-2 Harris simulation." Figure 3 is a boxological sketch of type-2 Harris simulation.

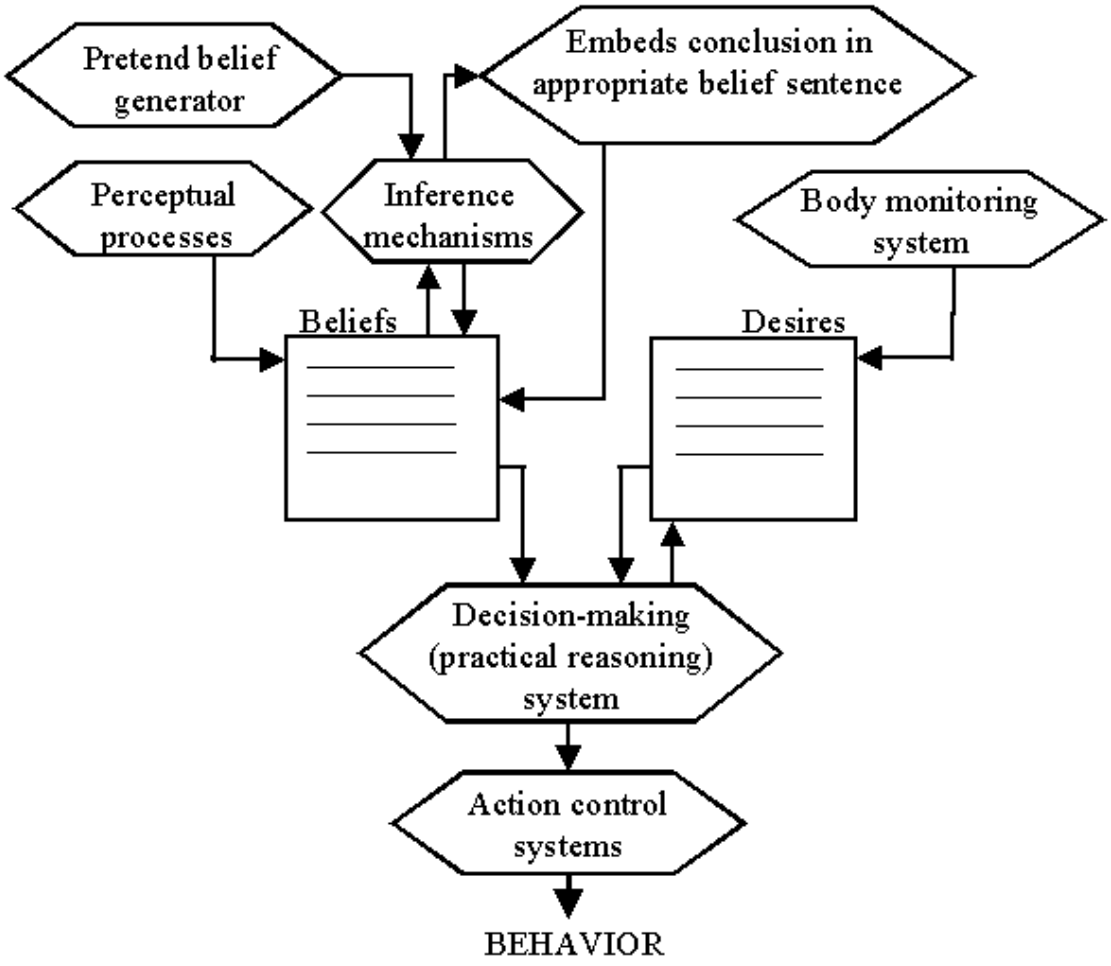

Figure 3

It is important to note that type-2 Harris simulation, like type-1 Harris simulation, is not a mechanism for predicting behavior. Rather, like type-1 Harris simulation, it is a mechanism for determining what someone else does or would believe. As was the case with type-1 Harris simulation, type-2 simulation is compatible with both the theory-theory account of folk psychological behavior prediction and with the off-line simulation account. The belief about what Sven believes might be used as a premise in a bit of reasoning that exploits the principles of folk psychology. Or it might be fed into the practical reasoning system, which would then produce an "off-line" decision about what to say.

We find the idea of type-2 Harris simulation to be quite intriguing. It may well be that some (or most) cases in which we form beliefs about what someone would infer proceed in this way. But it might also be that in some (or most) cases we use a tacit theory about how people go about the process of reasoning. To determine which of these hypotheses is correct, some careful experiments are needed to explore the extent to which reasoning about other people's reasoning is cognitively penetrable. On the type-2 Harris account, a person's ignorance or misinformation about how people go about the business of reasoning should be irrelevant to the accuracy of the answers he gives in cases like those we have been considering. But if we typically exploit a tacit theory about how people go about the business of reasoning, then ignorance or misinformation should seriously degrade the accuracy of our answers. Until the relevant experiments are done, we propose to remain agnostic about type-2 Harris simulation.

It's time to sum up the results of this section. We have tried to distinguish three quite different simulation hypotheses:

i) Off-line simulation

ii) Type-1 Harris simulation

iii) Type-2 Harris simulation.

Of these three, type-1 Harris strikes us as by far the most plausible. Type-2 Harris might well turn out to be correct; we're offering no bets on that one. We remain deeply skeptical about off-line simulation, for some of the reasons set out

in FPSTT. However, several of the contributors to this volume have raised interesting objections to the arguments in FPSTT. So in the next section we'll set out our replies to some of these objections.

3. Some Responses to Our Critics

As we see it, there are four areas in which our critics have raised important objections to the arguments set out in FPSTT.

3.1 Inference Neglect and Other Developmental Findings

The first of these objections questions the robustness of "inference neglect" and some of the other findings cited in Section 5 of FPSTT; the objection also questions the extent to which those data favor the theory-theory over the off-line simulation theory. On this point we propose to pull in our horns a bit. We relied heavily on the results reported in Wimmer et. al. (1988). But as Goldman, Gopnik and Wellman, and Gordon all have noted, there are some serious problems with these results. So until more work is done, we are reluctant to put much weight on inference neglect. However, it is also the case that there have been many new and interesting results reported in the developmental psychology literature since we finished work on FPSTT. We are inclined to think that by and large these results favor the theory-theory. But Gopnik and Wellman have already done an excellent job of making the case for this claim in their contribution to this volume. We have little to add.

3.2. The Argument from Simplicity

The second area in which important objections have been raised is the debate over simplicity. We argued that neither theory-theorists nor off-line simulation theorists can gain much advantage by appealing to simplicity. But Goldman has offered an interesting new argument aimed at showing that off-line simulation is in fact the simpler theory. As Goldman rightly notes, the essential claims in our argument were as follows: The off-line simulation theory gets its data base (the information it needs about the human decision making system) for free, but it must also posit a highly specialized "control mechanism" which provides pretend inputs and takes the decision maker off-line at appropriate times. By contrast, the theory-theory needs a very substantial data base (the tenets of folk psychology) but it gets its control mechanism (the system needed to extract appropriate information from an internalized theory and apply it to particular cases) for free. Thus it appears that in terms of simplicity the two theories are roughly on a par. But on Goldman's view, the simulation theory gets both the data base and the control mechanism for free. Here is how he makes the point:

> What do [Stich & Nichols] mean by saying that a theory gets a system "for free"? Evidently they mean that the system in question must be posited in any case, no matter what side is chosen in the current dispute. I accept this construal for purposes of comparisons of parsimony. On that very criterion, however, the simulation theory also gets its crucial control mechanism for free, since the "off-line" capacity must be posited in any case to handle other uncontroversial cognitive abilities!
>
> Consider the activity of conditional planning. A group of us are trying to decide which restaurant to dine at this evening, and I am the designated driver.... Although our choice is yet to be made, I begin to plot which route I would take if we chose restaurant A.... To make these conditional plans, I "pretend" to have the goal of going to restaurant A, and deliberate on the best means to get there (shortest time, least traffic, or whatnot).... Since I haven't actually formed a goal or intention to drive to either restaurant, my decision-making system is operating off-line. The output from this process is not an actual decision to take the selected route but merely a conditional decision, viz., to take that route if we decide to go to that restaurant. Since this kind of deliberation must be accommodated by any theory, we are going to have to posit "off-line" decision making in any case. Thus, the simulation theory gets this control mechanism for free. The result is that the simulation theory gets both its "data base" and its control mechanism for free.... Hence the simulation theory is indeed more parsimonious.[(15)]

Now we are certainly prepared to admit that people do engage in what Goldman calls "conditional planning". In cases like the one Goldman describes, the driver might indeed come to some view about which route he would decide to take if he should want to go to restaurant A and/or about which route is best given the relevant circumstances and constraints. But how would the driver come to this view? Goldman claims that he would take his decision making system "off-line" and feed it the "pretend" goal of getting to restaurant A. However, it is our contention that this claim simply begs the question. For there are various alternative hypotheses about the cognitive processes employed in these

sorts of cases - hypotheses which do not include an off-line use of the driver's decision making system.

For example, it might be the case that the formation of "conditional plans" is carried out by predicting one's own behavior in counterfactual circumstances, and that both these predictions about one's own behavior and predictions about other people's behavior are accomplished by invoking a tacit folk psychological theory. If this is the mechanism that subserves the formation of "conditional plans" it provides an obvious explanation for those cases in which we do not actually do what we had conditionally planned to do, when the situation for which we were planning actually comes to pass. For if we are using a theory to determine what we would do if a certain situation arises, then if the theory is mistaken it is not surprising that the prediction might come out wrong. It is far less obvious how a simulation theorist would handle such cases, though no doubt a resourceful simulation theorist could produce some sort of explanation.

Yet another possible explanation of cases like the one that Goldman describes is that, having noted that he might want to go to restaurant A, the agent is simply deliberating about the best way to get there given the prevailing circumstances. The deliberative processes may be similar to the one that an engineer might pursue in figuring out the best way to build a bridge at a given location, while taking account of various constraints. There is no obvious reason why this sort of problem solving should involve either pretense (as in Goldman's account) or prediction (as in our previous hypothesis).

We won't pursue these various hypotheses, since it is no part of our current project to determine which of them (if any) is correct. What is crucial for the issue at hand is that there are various explanations of the phenomenon that Goldman calls "conditional planning" that do not posit "off-line capacity". Thus, contrary to Goldman's contention, it is not the case that "the simulation theory ... gets its crucial control mechanism for free, since the `off-line' capacity must be posited in any case to handle other uncontroversial cognitive capacities." In arguing that the simulation theory gets the control mechanism for free, Goldman is simply assuming without argument that conditional planning is subserved by an off-line simulation process rather than by one of the other processes we have sketched. Once this undefended assumption is noted, his argument for the greater simplicity of the simulation theory collapses.

3.3. Cognitive Penetrability

In FPSTT we argued that the theory-theory and the off-line simulation theory differed dramatically in their expectations about cognitive penetrability. According to the theory-theory, predictions about people's behavior are guided by a rich body of mentally represented information (or mis-information) about the ways in which psychological states are related to environmental stimuli, other psychological states, and behavioral events. If that information is wrong or incomplete in various areas, then we should expect the accuracy of predictions in those areas to decline. According to the off-line simulation theory, we generate predictions of people's behavior by running our own decision making system off-line. If we are ignorant about how people's minds work, or if we have mistaken views, this should not affect the accuracy of our predictions about how people will behave, since our views about how the mind works are not involved in generating the predictions. So if the off-line simulation theory is right, what we don't know won't hurt us -- predictions about people's behavior are "cognitively impenetrable."

If we understand them correctly, all the contributors to this volume who address the issue of cognitive penetrability agree with this way of setting out the implications of the two theories. Thus producing evidence in favor of or against cognitive penetrability should be one of the best ways of deciding which theory is more plausible. To the best of our knowledge, no one has yet run any experiments specifically designed to test the cognitive penetrability of folk psychological predictions. But in FPSTT we noted several examples of anecdotal evidence which, we thought, suggested that such predictions were indeed cognitively penetrable.

One of our examples turned on the belief perseverance phenomenon that has been explored by Ross and his colleagues. (16) In Ross's experiments, subjects are given a little test which indicates that they are unusually good (or unusually bad) at a certain task. About an hour later it is explained to them that the test results were bogus. But despite this, when the subjects are later asked whether they would be good or bad at the task, those who were duped into thinking they were unusually good say they would be good at it; those who were duped into thinking that they were unusually bad say they would be bad at it.

Now let us offer some anecdotal evidence. Both of us have on occasion described the Ross experiment to students who did not know the outcome, and asked the students to predict what the subjects would say. The predictions the students offered were more often wrong than right. The explanation we propose turns on the fact that most people are quite ignorant of the belief perseverance phenomenon; the facts about perseverance have not made their way into folk

psychology. Clearly, the theory-theory has no trouble explaining why most of our students predicted incorrectly. If the simulation theory is right, however, then the students made their prediction by first pretending they were in subject's shoes, and then letting their normal cognitive process run "off-line". But if that's really what they did, then since it is likely that they themselves would manifest belief perseverance if they were subjects in Ross's experiments, it seems that they should have predicted correctly. Since the simulation theory entails that such predictions are cognitively impenetrable, the students' ignorance about perseverance should have been irrelevant.

None of our critics were very impressed by this sort of argument. All of them claimed that as we set things up in our various examples, the predictors were likely to run their simulations on the wrong "pretend inputs," and thus it is not surprising that they make the wrong predictions. Here is how Harris makes the point:

> In general, we may distinguish between two different sources of error [in predictions produced by adults]. First, ... it is necessary for the simulator to feed in pretend inputs that match in the relevant particulars the situation facing the agent whose actions are to be predicted or explained. Predictive errors will occur if inappropriate pretend inputs are fed in. Second, any simulation process assumes that an actor's behavior is a faithful translation into action of a decision that is reached by the practical reasoning system. If that assumption is incorrect, the simulation will err.
>
> ....In the case of belief perseverance, subjects in the experiments received two distinct and successive pieces of information. At first they were given apparently veridical information about a particular trait or competence, and later they were given information that discredited that initial information. By contrast, anyone reading about such experiments and attempting to simulate their outcome is presented with a single integrated account of both the trait information and its disconfirmation. They can judge the trait information for what it is worth as soon as they read about it. Under these circumstances, a reader will find it difficult to reproduce the naive, unsuspecting commitment to the initial information that is entertained by participants in the experiments. (pp. 132-133)

We take this to be a perfectly reasonable challenge to the sort of anecdotal evidence suggested in FPSTT. It is also a challenge that admits of a straightforwardly empirical response. If Harris is right, then our students simulated incorrectly because there was no time lag between presentation of the "trait information" and presentation of the "disconfirmation". So presumably Harris would predict that if we presented the information in two distinct phases, separated by an hour or so, people would make the correct prediction. While we haven't run any carefully controlled studies of this sort, we have recently tried Harris's procedure on a handful of students. Most of them still got the wrong answer. Moreover, there are some studies reported in the literature in which subjects predicted people's behavior incorrectly even when exposed to situations exactly like those to which the targets of the prediction were exposed.

Perhaps the most striking of these is an experiment by Bierbrauer (1973) which reenacted Milgram's classic obedience experiment (Milgram, 1963) in order to study the predictions that would be made by non-participant observers.

> In Bierbrauer's study, observers were exposed to a verbatim reenactment of one subject's obedience to the point of delivering the maximum shock to the supposed victim, and then were asked to predict how other subjects would behave. Bierbrauer's observers consistently and dramatically underestimated the degree to which subjects generally would yield to the situational forces that they had viewed at first hand. The observers did so, moreover, even in a set of conditions in which they themselves played the role of a subject in a vivid reenactment of Milgram's experiment.[(17)]

Now, of course our informal experiment is not an adequate substitute for a carefully controlled study. And the details of Bierbrauer's experiment were not specifically designed to explore the extent that predictions are cognitively penetrable. But in the light of Bierbrauer's results and our own, we think it is overwhelmingly likely that when a careful study of penetrability is run, the subjects will frequently predict incorrectly.

One of the reasons we admire Harris's critique of our penetrability argument is that it suggests a relatively straightforward experimental test that will go a long way toward deciding which of the two theories is correct. However, in discussing these matters with other defenders of the simulation theory we've come to suspect that some of them are not as willing as Harris is to subject their view to experimental disconfirmation. "Pretend" input to the decision making system is always going to be different from real input in some way or other. Thus no matter what experiment is proposed, if the result of the experiment seems to indicate that predictions are cognitively penetrable, it will always be possible for the resolute defender of the simulation theory to point to some such difference and protest

that the predictor hasn't used the right "pretend" input. But as we see it, this strategy is doubly problematic. First, and most obvious, it tends to make the cognitive impenetrability thesis empirically untestable. Unless the advocates of the simulation theory are prepared to specify in advance which features "of the situation facing the agent whose actions are to be predicted" are "relevant" and thus must be "matched" by the "pretend inputs" used by the simulator, no experimental disconfirmation of cognitive impenetrability will be possible. Second, it leaves the simulation theorist with a puzzle. Since "pretend" inputs are never exactly the same as real ones, why is it that folk psychological prediction is usually right? We don't propose to dwell on these matters, however, since we think the virtues of Harris's forthright approach are very clear. If the sorts of experiments that Harris's critique suggests come out as we expect they will, then the simulation theory should be abandoned; if they come out the other way, then it is the theory-theory that is in deep trouble.

3.4. Autism, Empathy and Understanding Mental States

Autistic people manifest an array of symptoms which have lead a number of investigators to conclude that they lack a fully developed theory of the mind. But Goldman offers several considerations which, he contends, support a rather different conclusion. According to Goldman, the data are better explained by the hypothesis that autistic people suffer from a "deficiency in simulation."

One of Goldman's arguments turns on the fact that autistic people seem incapable of empathizing with other people.

> [A]utistic people are noted for their indifference to other people's distress and their inability to offer comfort, even to receive comfort themselves (see Frith, 1989, pp. 154-5). This obviously could be associated with a deficiency in simulative propensities.(18)

Now it strikes us that all this is a bit fast and loose. It is not at all clear how a "deficiency in simulative propensities" would explain the fact that autistic people are indifferent to other people's distress. There are, after all, lots of states of affairs about which we are not at all indifferent, but which we haven't any idea how to simulate. We are both deeply distressed by the destruction of the rain forest, for example, though we would hardly know what it meant to "simulate" the destruction of the rain forest. So, in this case at least, inability to simulate is quite compatible with caring deeply. Still, perhaps there is some way in which "a deficiency in simulative propensity" might be used to explain the indifference to distress observed in autistic people. If so, a detailed statement of how the explanation is supposed to run would be a welcome contribution to the debate. But even if such an explanation can be given, it is far from clear that it would count as an argument in favor of the simulation theory and against the theory-theory. For, as Frith herself has noted, the more standard view according to which autistic people have a defective theory of mind provides an obvious explanation for the phenomenon. If autistic people's mastery of the theory of mind is so fragmentary that they don't even realize that other people are in distress, it is hardly surprising that they will not empathize with other people's distress.

> [I]f autistic people cannot conceptualize mental states very well, then they cannot empathize with the mental states of others, such as their feelings.(19)

A second argument offered by Goldman turns on a pair of anecdotes in which autistic people misunderstand requests in quite striking ways. In one of these anecdotes, while baking a mother said to her autistic son, "I haven't got any cloves. Would you please go out and get me some?" The son came back with a bag full of woman's clothes from a High Street boutique. According to Goldman, this anecdote indicates that there is nothing wrong with the young man's understanding of the mentalistic notion of wanting. The defect is in his capacity to simulate.

> [I]t seems to me that this autistic young man has a perfectly good grasp of the mentalistic notion of wanting. He understood that his mother wanted something, and her (misperceived) utterance was interpreted as stemming from that want. No deficiency in the concepts or relations of "theory of mind" appear here.... What is striking is the son's attribution to his mother of an outlandish desire, which anybody else would reject immediately. This is naturally explainable as a failure to project himself into her shoes in an adequate way, and thereby eliminate the hypothesis of this particular desire. In short, it is a deficiency in simulation.(20)

Now, as we see it, the problem with Goldman's interpretation of the anecdote comes right at the beginning. Neither the term "want" nor any other mentalistic language occurs in the anecdote, and there is no indication that the young man "has a perfectly good grasp of the mentalistic notion of wanting." The mother issues a request, and it is entirely possible

that the autistic son understood the request "behavioristically" -- when mother (or some other person of authority) issues a request, one does what is requested. Much the same interpretation is possible for the other anecdote Goldman recounts. In that one, an autistic girl showed great anxiety when a nurse about to do a blood test said, "Give me your hand." Here again, no mentalistic language is used. There is no evidence that the girl attributed a ghoulish desire to the nurse, or that she attributed any desire at all. Once again, she may simply have interpreted the command behavioristically and become frantic when she took the command literally.

While we don't think that Goldman's anecdotes provide any reason at all to favor the simulation theory over the theory-theory, we do think they are interesting for quite a different reason. For Goldman's explanation of the behavior recounted in the anecdotes suggests yet another variation on the simulation theme. The sort of off-line simulation sketched in Figure 2 uses the "practical reasoning" system to produce decisions about what to do. Since the system is running off-line, these decisions are not acted on. Instead they are transformed into predictions about the behavior of the person being simulated. But, of course, when running on-line the practical reasoning system has another function. It produces new desires. If an agent wants to follow the recipe in certain cookbook, and if she comes to believe that the recipe requires cloves, then (typically) the agent will form a desire to have some cloves. Sometimes these lower level or "instrumental" desires lead to decisions which are then acted on. But in many cases they don't, since the agent recognizes that, given her circumstances there is nothing much she can do to satisfy the desire, at least for the moment. So the desire is simply stored, to be acted upon when the opportunity arises. Now suppose this entire process can be taken off-line. We feed our practical reasoning system some pretend beliefs and/or pretend desires which we have reason to think that some other person (the "target") has. We then let the practical reasoner generate the appropriate instrumental desires. These don't become our desires, however. Rather, information about the content of the desires is embedded in appropriate desire sentences, which attribute the desires to the target. These desire sentences are then stored in our "belief box". In this way, we might use simulation to form beliefs about what other people desire. For want of a better label, we will call this process "off-line desire simulation." Figure 4 is a boxological rendition of the process. Whether or not people actually use this process to form beliefs about other people's desires is, of course, very much an open question. One way explore the question is to determine whether the process of forming beliefs about other people's desires is cognitively penetrable. Our guess is that the process is cognitively penetrable, and thus that off-line desire simulation is not the process people use. But the issue is obviously an empirical one, and a well designed experiment will go a long way toward settling it.

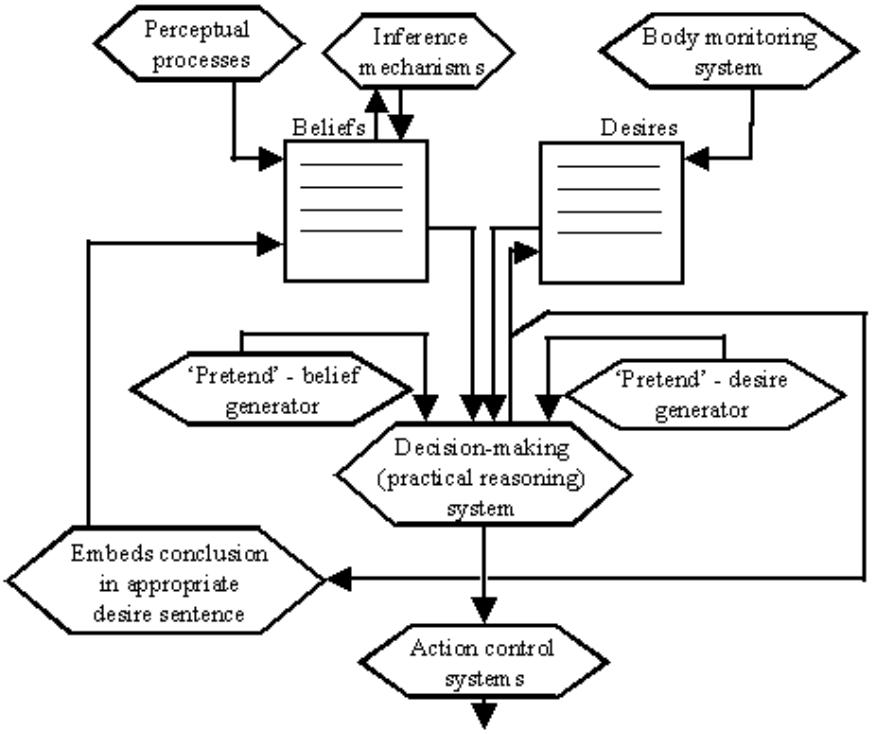

Figure 4

---

4. Conclusion

In FPSTT we claimed that there were no good arguments in favor of the off-line simulation theory, and several strong arguments against it. A substantial part of the current paper has been devoted to responding to objections raised by our critics. As we see it, the bottom line on off-line simulation remains unchanged: there are still no good arguments in favor of the theory, and there are still strong arguments against it. In the process of replying to our critics we have had occasion to distinguish various different versions of the theory-theory and various different versions of the simulation theory. We are inclined to think that these distinctions are both more interesting and more valuable than the various polemical points we've tried to score. For they indicate that the dispute between theory-theorists and simulation theorists is much more complex than has hitherto been recognized. It may well turn out that some of our folk psychological skills are indeed subserved by simulation processes, while others are subserved by processes that exploit a tacit theory. Whether this is the case, and, if so, which skills rely on which processes, are matters that can only be settled by doing the appropriate experiments. It looks like there is lots of work to be done before we have a good understanding of how people go about attributing mental states to each other and predicting each other's behavior.[(21)]

REFERENCES

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

Fodor, J. (1983). *The Modularity of Mind*, Cambridge, MA, MIT Press.

Fodor, J. (1987). *Psychosemantics*, Cambridge, MA, MIT Press.

Frith, U. (1989). *Autism: Explaining the Enigma*, Oxford, Basil Blackwell.

Goldman, A. (1992). In Defense of the Simulation Theory," this volume.

Gordon, R. (1992). "Reply to Stich and Nichols," this volume.

Gopnik, A. and Wellman, H. (1992). "Why the Child's Theory of Mind Really Is a Theory," this volume.

Johnson-Laird, P. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference and Consciousness*, Cambridge, MA, Harvard University Press.

Lewis, D. (1972). "Psychophysical and Theoretical Identifications," *Australasian Journal of Philosophy*, 50, 249-258. Reprinted in Ned Block, ed., *Readings in the Philosophy of Psychology*, Volume 1, Harvard University Press, 1980.

Milgram, S. (1963). "Behavioral Study of Obedience," *Journal of Abnormal and Social Psychology*, 67.

Nichols, S., Stich, S., Leslie, A., and Klein, D. 1996. "Varieties of Off-Line Simulation". In *Theories of Theories of Mind*, eds. P. Carruthers and P. Smith. Cambridge: Cambridge University Press, 39-74.

Nisbett, R. and Ross, L. (1980). *Human Inference: Strategies and Shortcomings of Social Judgement*, Englewood-Cliffs, NJ, Prentice-Hall.

Ross, L., Lepper, M., and Hubbard, M. (1975). "Perseverance in Self-Perception and Social Perception: Biased Attributional Processes in the Debriefing Paradigm," *Journal of Personality and Social Psychology*, 32.

Stich, S. (1978). "Beliefs and Subdoxastic States," *Philosophy of Science*, 45, 499-518.

Stich, S. and Nichols, S. (1992). "Folk Psychology: Simulation or Tacit Theory?" *Mind and Language*, 7, 1&2. Reprinted in this volume.

Wellman, H. (1990). *The Child's Theory of Mind*, Cambridge, MA, MIT Press.

Wimmer, H., Hogrefe, J. and Sodian, B. 1988. A Second Stage in Children's Conception of Mental Life: Understanding Informational Access as Origins of Knowledge and Belief. In Astington, J., Harris, P. and Olson, D. (eds.), *Developing Theories of Mind*, Cambridge, Cambridge University Press, 173-192.

NOTES

1. Though one of the boxes in Figure 1 is labeled "BELIEFS" something like "BELIEFS AND SUBDOXASTIC STATES" might be less misleading label. We are not assuming that all of the information stored in the belief box is accessible to consciousness. Nor do we assume that everything in the box is inferentially integrated with everything else; there may well be sub-components (or "modules") of the belief box that are largely inferentially isolated from the rest of the system. Finally, we do not assume that everything in the box is stored in the same format. Parts of the system may be stored sententially, other parts may be stored pictorially, and still other parts may be stored in other ways. For more on the idea of a "subdoxastic state" see Stich (1978); for further discussion of cognitive "modules" see Fodor (1983).

2. "Folk Psychology: Simulation or Tacit Theory?" this volume. Hereafter we'll refer to this paper as FPSTT.

3. See, for example, Johnson-Laird (1983).

4. Those who assume that folks psychology will include lawlike generalizations include Fodor (1987), Wellman (1990), Goldman and Gordon.

5. The quote is from Gopnik and Wellman, this volume, p. 147. (Page reference is to *Mind and Language*, 7, 1992.) See also Wellman (1990), pp. 6-11 & p. 325.

6. "In Defense of the Simulation Theory," this volume, p. 106. (Page reference is to *Mind and Language*, 7, 1992.)

7. Lewis (1972).

8. "In Defense of the Simulation Theory,." p. 106. (Page reference is to *Mind and Language*, 7, 1992.)

9. "In Defense of the Simulation Theory,." p. 107. (Page reference is to *Mind and Language*, 7, 1992.)

10. Figure 2 in this paper is the same as Figure 3 in FPSTT.

11. "Reply to Stich and Nichols," this volume, p. 87. (Page reference is to *Mind and Language*, 7, 1992.)

12. See FPSTT, Section 4, Argument 5.

13. "From Simulation to Folk Psychology: The Case for Development," this volume. (Page references are to *Mind and Language*, 7, 1992.

14. Ibid., p. 124.

15. Ibid., pp. 110 - 111.

16. See, for example, Ross, Lepper and Hubbard (1975). For an excellent review of the literature, see Nisbett and Ross (1980).

17. Nisbett & Ross (1980), p. 121.

18. "In Defense of the Simulation Theory," p. 113. (Page reference is to *Mind and Language*, 7, 1992.)

19. Frith (1989), p. 167.

20. "In Defense of the Simulation Theory," ms. p. 112. (Page reference is to *Mind and Language*, 7, 1992.)

21. Much of the work on this paper was done while one of us, Stephen Stich, was a Visiting Fellow at the Research School of Social Sciences at the Australian National University. He would like to thank the RSSS, Prof. Frank Jackson and Dr. Karen Neander for their hospitality. Thanks are also due to Joseph Franchi for help in preparing the Figures. The paper was completed in November 1992. Since then our views have continued to evolve. See Nichols et al. 1996.