

And (Gordon, this volume, p. 72 n. 4):

Imagery is not always needed in such simulations. For example, I need no imagery to simulate having a million dollars in the bank.

To deal with this 'challenge', Goldman proposes a pair of replies. First, simulation need not always be introspectively vivid. It can often be 'semi-automatic, with relatively little salient phenomenology'. Second, not all interpretations rely on simulation. In many cases interpreters rely solely on 'inductively acquired informations' though the information is 'historically derived from earlier simulations' (this volume, pp. 87–8).

Reply: We don't propose to make any fuss at all about the frequent absence of 'salient phenomenology'. For it is our contention that when the issue at hand is the nature of the cognitive mechanism subserving our capacity to interpret and predict other people's behavior, the entire issue of introspective imagination is a red herring. Indeed, it is *two* red herrings. To see the first of them, consider one of the standard examples used to illustrate the role of imagery in thought. Suppose we ask you: 'How many windows are there in your house?' How do you go about answering? Almost everyone reports that they *imagine* themselves walking from room to room, counting the windows as they go. What follows from this about the cognitive mechanism that they are exploiting? Well, one thing that surely *does not* follow is that off-line simulation is involved. The *only* way that people could possibly answer the question accurately is to tap into some internally represented store of knowledge about their house; it simply makes no sense to suppose that off-line simulation is being used here. So even if a cognitive process is *always* accompanied by vivid imagery, that is no reason at all to suppose that the process exploits off-line simulation. From this we draw the obvious conclusion. The fact that prediction and interpretation *sometimes* involve imagining oneself in the other person's shoes is less than no reason at all to suppose that off-line simulation is involved.

It might be suggested that, though imagery provides no support for the off-line simulation hypothesis, it does challenge the theory-theory when 'theory' is interpreted narrowly. For it shows that some of the information we are exploiting in interpretation and prediction is not stored in the form of sentences or rules. But even this is far from obvious. There is a lively debate in the imagery literature in which 'descriptonalists', like Pylyshyn and Dennett, maintain that the mechanisms underlying mental imagery exploit language-like representations, while 'pictorialists', like Kosslyn and Fodor, argue that images are subserved by a separate, non-linguistic sort of representation (Pylyshyn, 1981; Dennett 1969, 1978b; Fodor, 1975; Kosslyn, 1981). We don't propose to take sides in this dispute. For present purposes it is sufficient to note that, unless it is supplemented by a persuasive argument in favor of pictorialism and against descriptionism, the introspective evidence does not even challenge the theory-theory *construed narrowly*.

Argument 6: The off-line simulation account is supported by recent experimental studies focusing on children's acquisition of the ability to interpret and predict other people.

On our view, this is far and away the most interesting argument that has been offered in favor of the off-line simulation theory. To see exactly what the experimental studies do, and do not, support, we'll have to look at both the evidence and the argument with considerable care. Gordon does a good job of describing one important set of experiments (this volume, pp. 68–9):

Very young children give verbal expression to predictions and explanations of the behavior of others. Yet up to about the age of four they evidently lack the concept of belief, or at least the capacity to make allowances for false or differing beliefs. Evidence of this can be teased out by presenting children with stories and dramatizations that involve dramatic irony: where we the audience know something important the protagonist doesn't know . . .

In one such story (illustrated with puppets) the puppet-child Maxi puts his chocolate in the box and goes out to play. While he is out, his mother transfers the chocolate to the cupboard. Where will Maxi look for the chocolate when he comes back? In the box, says the five-year-old, pointing to the miniature box on the puppet stage: a good prediction of a sort we ordinarily take for granted . . . But the child of three to four years has a different response: verbally or by pointing, the child indicates the cupboard. (That is, after all, where the chocolate is to be found, isn't it?) Suppose Maxi wants to mislead his gluttonous big brother to the *wrong* place, where will he lead him? The five-year-old indicates the cupboard, where (unbeknownst to Maxi) the chocolate actually is . . . The *younger* child indicates, incorrectly, the box.

These results, Gordon maintains, are hard to square with the theory-theory. For if the theory-theory is correct, then (pp. 69–70):

before internalizing [the laws and generalizations of folk psychology] the child would simply be unable to predict or explain human action. And *after* internalizing the system, the child could deal indifferently with actions caused by *true* beliefs and actions caused by *false* beliefs. If is hard to see how the semantical question could be relevant.

But, according to Gordon, these data are just what we should expect, if the off-line simulation theory is correct (Gordon, unpublished, sec. 3.6, p. 11):

The Simulation Theory [predicts that] prior to developing the capacity to simulate others for purposes of prediction and explanation, a child will make *egocentric errors* in predicting and explaining the actions of others. She will predict and explain as if whatever she herself counts as

'fact' were also fact to the other; which is to say, she fails to make allowances in her predictions and explanations for false beliefs or for what the other isn't in a position to know.

Reply: According to Gordon, the theory-theory can't easily explain the results of the 'Maxi' experiment, though the off-line simulation theory predicts those results. We're not convinced on either score. Let's look first at just what the off-line simulation story would lead us to expect.

Presumably by the time any of these experiments can be conducted, the child has developed a more or less intact decision-making system like the one depicted in figure 5.1. That system makes 'on-line' decisions and thus determines the child's actions on the basis of her actual beliefs and desires. But by itself it provides the child with no way of predicting Maxi's behavior or anyone else's. If the off-line simulation theory is right, then in order to make predictions about other people's behavior two things must happen. First, the child must acquire the ability to take the output of the decision-making system off-line – treating its decisions as predictions or expectations, rather than simply feeding them into the action controlling system. Second, the child must acquire the ability to provide the system with input other than her own actual beliefs and desires. She must be able to supply the system with 'pretend' input so that she can predict the behavior of someone whose beliefs and desires are different from her own. (These are the two capacities that are represented in figure 5.3 and absent in figure 5.1.) There is, of course, no a priori reason to suppose that these two steps happen at different times, nor that the one we've listed first will occur first. But if they do occur in that order, then we might expect there to be a period when the child could predict her own behavior (or the behavior of someone whose beliefs and desires are the same as hers), though she could not predict the behavior of people whose beliefs or desires are different from hers. It is less clear what to expect if the steps occur in the opposite order. Perhaps the result would be some sort of pretending or play-acting – behaving in a way that someone with different beliefs or desires would behave. Though until the child develops the capacity to take the output of the decision-making system off-line, she will not be able to predict other people's behavior or her own. So it looks like the off-line simulation story makes room for three possible developmental scenarios.

- (1) The child acquires both abilities at the same time. In this case we would expect to see two developmental stages. In the first the child can make no predictions. In the second she can make a full range of predictions about people whose beliefs and desires are different from her own.
- (2) The child first acquires the ability to take the output off-line, and then acquires the ability to provide the system with pretend input. In this case we would expect three developmental stages. In the first, the child can make no predictions. In the second, she can only make predictions about her own behavior or about the behavior of people whose beliefs and

desires are identical to hers. In the third, she can make the full range of predictions.

- (3) The child first acquires the ability to provide the system with pretend inputs, and then acquires the ability to take the output off-line. In this case, too, we would expect three developmental stages. The first and last stages are the same as those in (2), but in the middle stage the child can play-act but not make predictions.

Now let's return to the Maxi experiment. Which of these developmental scenarios do the children in these experiments exhibit? At first blush, it might be thought that the pattern Gordon reports is much the same as the one set out in scenario (2). But that would be a mistake. The younger children – those who are giving the wrong answers – are not predicting that Maxi would do what someone with their own beliefs and desires would do. For they have no desire to get the chocolate, nor to deceive the gluttonous brother. Those are *Maxi's* desires, not *theirs*. If anything, it would appear that these children are halfway between the second and third stages of scenario (2): they can feed 'pretend' desires into the decision-making system, but not 'pretend' beliefs. Of course none of this shows that the off-line simulation theory is false. It is perfectly compatible with the theory to suppose that development proceeds as in (2), and that the transition from the second to the third stage proceeds in two sub-stages – desires first, and then beliefs. (This pattern is sketched in figure 5.4.) But it is, to say the least, something of an exaggeration to say that the off-line simulation theory 'predicts' the experimental results. The most that can be said is that the theory is compatible with the observed developmental pattern, and with lots of other patterns as well.⁹

For the results that Gordon describes to be at all relevant to the dispute between the off-line simulation theory and the theory-theory, it would have to be the case that the latter theory is *incompatible* with the reported developmental pattern. But that is patently not the case. To see why, we should first note that the theory-theory is not committed to the claim that folk psychology is acquired all in one fell swoop. Indeed, one would expect just the opposite. If children really are acquiring a tacit theory of the mind, they probably acquire it a bit at a time. Thus it might be the case that, at a given stage in development, children have mastered the part of the theory that specifies how beliefs and desires lead to behavior, though they have not mastered the entire story about how beliefs are caused. At this stage, they might simply assume that beliefs are caused by the way the world is; they might adopt the strategy of attributing to everyone the very same beliefs that they have. A child who has acquired this much of folk psychology would (incorrectly) attribute to Maxi the belief that the chocolate is in the cupboard. She would then go on to make just the predictions that Gordon reports. Of course, the theory-theory is also compatible with lots of other hypotheses about which bits of folk psychology are acquired first. Thus, like the off-line simulation theory, it is compatible with (but does not entail) lots of possible developmental patterns.

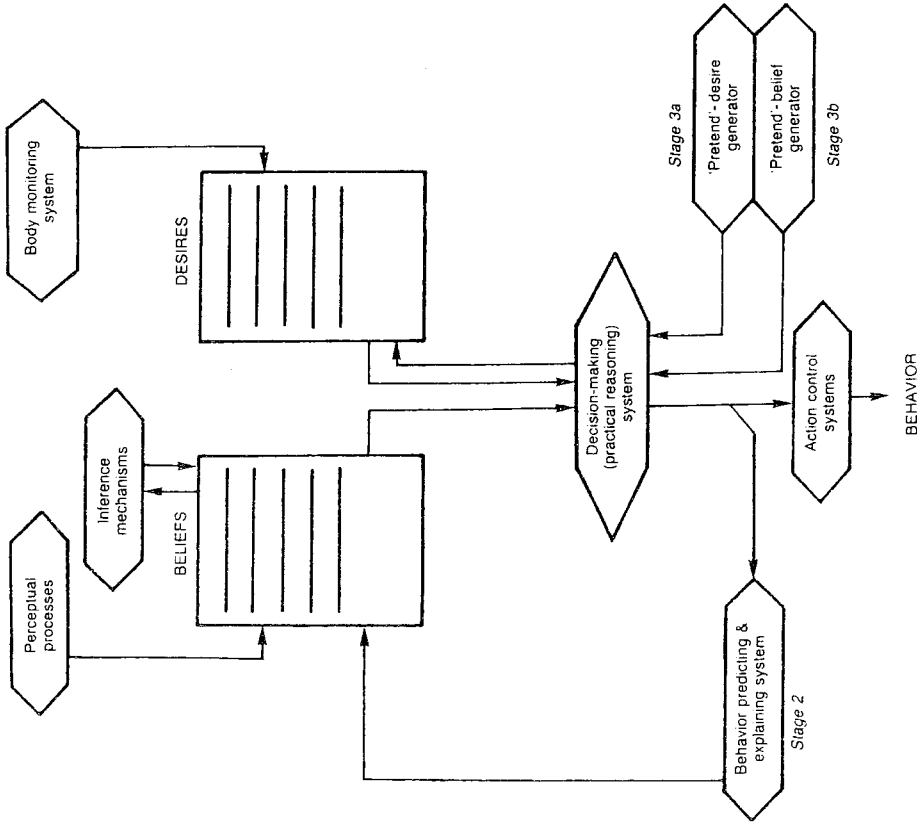


Figure 5.4

So it looks like the developmental studies that Gordon and Goldman cite can't be used to support one theory over the other.

Argument 7: Autistic children are highly deficient in their ability to engage in pretend play. These children are also frequently unable to impute beliefs to others or to predict other people's behavior correctly.

Here's how Gordon sets out the argument (this volume, p. 70):

Practical simulation involves the capacity for a certain kind of systematic pretending. It is well known that *autistic* children suffer a striking

deficit in the capacity for pretend play. In addition, they are often said to 'treat people and objects alike'; they fail to treat others as subjects, as having 'points of view' distinct from their own. This failure is confirmed by their performance in prediction tests like the [Wimmer-Perner 'Maxi' experiment] I have just described. A version of the Wimmer-Perner test was administered to autistic children of ages six to sixteen by a team of psychologists . . . *Almost all* these children gave the wrong answer, the 3-year-old's answer. This indicates a highly specific deficit, not one in general intelligence. Although many autistic children are also mentally retarded, those tested were mostly in the average or borderline IQ range. Yet children with Down's syndrome, with IQ levels substantially below that range, suffered no deficit: almost all gave the right answer. My account of belief would predict that only those children who can engage in pretend play can master the concept of belief.

Goldman is rather more tentative. He claims only that the inability of autistic children 'to impute beliefs to others and therefore predict their behavior correctly . . . might . . . be related to their lack of pretend play' (this volume, p. 87).

Reply: The fact that autistic children are both incapable of pretend play and unable to predict the behavior of other people in Wimmer-Perner tests is very intriguing. Moreover, Gordon is certainly right in suggesting that the off-line simulation theory provides a possible explanation for these facts. If the off-line simulation theory is right, predicting the behavior of people whose beliefs differ from our own requires an ability to provide our own decision-making system with pretend input. And it is plausible to assume that this ability would also play a central role in pretend play. So if we hypothesize that autistic children lack the ability to provide the decision-making system with pretend input, we could explain both their performance on the Wimmer-Perner test and their failure to engage in pretend play. But, of course, this will not count as an argument for the off-line simulation theory and against the theory-theory if the latter account can offer an equally plausible explanation of the facts. And it will require no creativity on our part to produce such an alternative explanation since one of the investigators who discovered the fact that autistic children do poorly in Wimmer-Perner tests has offered one himself.

Leslie (1988) takes as an assumption 'the hypothesis that human cognition involves *symbolic computations* in the sense discussed . . . by Newell (1980) and particularly by Fodor' (Leslie, 1988, p. 21). He also assumes that an internalized theory of mind underlies the normal adult's ability to predict other people's behavior. An important theme in Leslie's work is that developmental studies with both normal and autistic children can help to illuminate the expressive resources of the 'language of thought' in which our theory of mind is encoded. According to Leslie, the notion of a 'meta-representation' is central in understanding how our theory of mind develops. Roughly speak-

ing, a meta-representation is a mental representation about some other representational state or process. We exploit meta-representations when we think that

Maxi believes that the chocolate is in the box
or that

Maxi's brother wants the chocolate
or that

Mommy is pretending that the banana is a telephone.

On Leslie's view, 'autistic children do not develop a theory of mind normally' (Leslie, 1988, p. 39). And while 'it is far too soon to say with any confidence what is wrong' with these children, he speculates that at the root of the problem may be an inability to use meta-representations. If this were true, it would explain both their difficulty with pretend play and their failure on the Wimmer-Perner test.

Though we find Leslie's speculation interesting and important, it is no part of our current project to defend it. To make our case we need only insist that, on currently available evidence, Leslie's hypothesis is no less plausible than Gordon's. Since Leslie's speculation presupposes that normal children acquire and exploit a theory of mind that is encoded in a language of thought, the evidence from studies of autistic children gives us no reason to prefer the off-line simulation account over the theory-theory.¹⁰

Our theme, in this reply and in the previous one, has been that the empirical evidence cited by Gordon and Goldman, while compatible with the off-line simulation theory, is also compatible with the theory-theory, and thus does not support one theory over the other. But there are other studies in the recent literature that *can* be used to support one theory over the other. These studies report results that are comfortably compatible with the theory-theory though not with the off-line simulation account. Before we sketch those results, however, it is time to start a new section. In this section we've tried to show that none of the arguments in favor of the off-line simulation theory is persuasive. In the next one we'll set out a positive case for the theory-theory.

5 *In Defense of the Theory-Theory*

Argument 1: There are developmental data that are easily accommodated by the theory-theory, but very hard to explain if the off-line simulation account is correct.

Let's start with a description of the experimental setup, and a quick overview of the data.

The setup of the task in these experiments was rather simple. Two children were placed facing each other on opposite sides of a table. In each trial one child served as subject and had access to the other child's knowledge and his or her own knowledge of the content of a closed box. The box was placed in the middle of the table between the two children. The outside of it was neutral and not suggestive of its content. In each box was a familiar object like a pencil, a comb, a piece of chocolate, and so on. The specific questions were: 'Does (name of other child) know what is in the box or does she not know that?' and 'Do you know what is in the box or do you not know that?' ...

Before the knowledge-questions were asked, either the other or the subject had access to the content of the box. One kind of access was visual perception. In this case either the other child or the subject had a chance to look into the box. The other kind of access was verbal information. Here the experimenter looked into the box and then informed one of the children by whispering the name of the content object into the child's ear. Because the two children were facing each other the subject was fully aware of the information conditions the other child was exposed to, that is, of whether the other child did or did not look into the box and of whether the other was or was not informed. (Wimmer et al., 1988, p. 175)

The results of this experiment were quite striking. The older children (five-year-olds) gave uniformly correct answers. But younger children (three- and four-year-olds) did not.

The most frequent error was denial of the other child's knowledge when the other child had looked into the box or was informed by the experimenter.

Most 3-year-olds and some 4-year-olds said that the other did not know what was in the box. This kind of error was nearly absent in children's assessment of their own knowledge. When subjects themselves had looked into the box or were informed, then they claimed to know and they could, of course, tell what was in the box. (Wimmer et al., 1988, pp. 175–6)

In another experiment, designed to be sure that the younger children were aware the other child had looked in the box, the subjects were asked both whether the other child had looked in the box and whether the other child knew what was in the box. 'The children consistently responded affirmatively to the look-question but again quite frequently responded negatively to the knowledge question' (Wimmer et al., 1988, p. 176).

What is going on here? The explanation offered by the experimenters is that younger children are using quite different mental processes in assessing what they know and in assessing what the other child knows. To answer the question, 'Do you know what is in the box?' the children use what the

experimenters call the 'answer check procedure'. They simply check to see whether they have an answer to the embedded question in their knowledge base, and if they do they respond affirmatively. To answer the question about the other child's knowledge, the older children used what the experimenters call a 'direct access check procedure'. In effect, they ask themselves whether the other child looked in the box or was told about its contents. If so, they respond affirmatively. If not, they respond negatively. However, the three-year-olds did not use this procedure. They simply checked whether the other child had uttered a correct statement about the box's contents. If she had not, the subject said the other did not know. A very natural way to describe the situation is that while the younger children know that people who say *that p* typically believe or know *that p*, these children have not yet learned that people will come to know *that p* by seeing or being told *that p*. The younger children have acquired a fragment of folk psychology, while the older children have acquired a more substantial piece of the theory.¹¹ The older children have not, however, entirely mastered the theory, as indicated by another series of experiments.

These experiments focused on the role of *inference* in the acquisition of knowledge or belief. What they show is that 'four- and five-year olds relied on inference in their own acquisition of knowledge but denied that the other person might know via inference' (Wimmer et al., 1988, p. 179):

Inferential access was realized in these experiments in a very simple and concrete way. In a first step the child and the other person together inspected the content of a container and agreed that only sweets of a certain kind, for example, black chocolate nuts, were in the container. In a second step the other person or the subject was prevented from seeing how one choconut was transferred from the container into an opaque bag. However, this person was explicitly informed by the experimenter about this transfer, for example, 'I've just taken one of the things out of this box and put it in the bag'.

The condition where knowledge could be acquired via simple inference was contrasted with a condition where knowledge depended on actually seeing the critical object's transfer. In this latter condition two kinds of sweets were in the original container, and thus one could only know what the content of the critical bag was by having seen the transfer from container to bag.

Once again, the results were quite striking. In most cases the older children (six-year-olds, in this case) generally gave the right answers both about their own knowledge and about the other child's knowledge. But although the four-year-olds used inference in forming their own beliefs, a substantial majority of them exhibited a pattern that the experimenters called 'inference neglect'.

The response pattern 'inference neglect' means that the other person was assessed according to perceptual access: When the other person

saw the object's transfer to the bag, 4-year-olds attributed knowledge; when the other did not see this transfer, ignorance was attributed even when the other person in fact knew via inference. (Wimmer et al., 1988, p. 179)

One plausible way of accounting for these results is to hypothesize that the older children had mastered yet another part of the adult folk psychology. They had learned that knowledge and beliefs can be caused by inference as well as by direct perceptual access. And, indeed, this is just the interpretation that the experimenters suggest.

In contrast to the 3-year-olds discussed in the previous [experiment], the 4-year-olds and 5-year-olds in the present experiments understood quite well that one has to consider the other person's informational conditions when one is questioned about the other person's knowledge. Their only problem was their limited understanding of informational conditions. They understood only direct visual access as a source of knowledge and this led them to mistaken but systematic ignorance attributions in the case of inferential access. (Wimmer et al., 1988, p. 181)

Let's now ask what conclusions can be drawn from these experiments that will be relevant to the choice between the off-line simulation theory and the theory-theory. A first obvious fact is that the data are all comfortably compatible with the theory-theory. Indeed, the explanation of the data offered by the experimenters is one that presupposes the correctness of the theory-theory. What appears to be happening is that as children get older, they master more and more of the principles of folk psychology. By itself, of course, the theory-theory would not enable us to predict the data, since the theory-theory does not tell us anything about the order in which the principles of folk psychology are acquired. But the pattern of results described certainly poses no problem for the theory-theory.

The same cannot be said for the off-line simulation theory. It is clear that even the younger children in these studies form beliefs as the result of perception, verbally provided information, and inference. So there is nothing about their decision making system, when it is being used on-line, that will help to explain the results. To make predictions about other people, the off-line simulation theory maintains, children must acquire the capacity to take the decision-making system off-line and provide it with some pretend inputs. But there is no obvious way in which this process could produce the pattern of results that has been reported. The difficulty is particularly clear in the case of inference. If the subject has seen that the box contains only chocolate nuts, and if she is told that one of the items in the box has been put in the bag, she comes to believe that there is a chocolate nut in the bag. But if she knows the other child has also seen what is in the box, and that the other child has been told that one of the items in the box has been put in the bag, she insists that the other child does not know what is in the bag. The problem can't be that the subject doesn't realize that the other child knows what is in the box. Children

of this age do a good job of attributing belief on the basis of perception. Nor can it be that the subject doesn't believe that the other child believes the transfer has been made. For children of this age are also adept at attributing beliefs on the basis of verbally communicated information. So it looks like the subject has all the information needed for a successful simulation. But the answer she comes up with is *not* the one that she herself would come up with, were she in the subject's place. There are, of course, endlessly many ways in which a resolute defender of the off-line simulation theory might try to accommodate these data. But all the ones we've been able to think of are obviously implausible and *ad hoc*.

Argument 2: Our predictions and explanations of behavior are 'cognitively penetrable'.

One virtue of using a simulation to predict the behavior of a system is that you need have no serious idea about the principles governing the behavior of the target system. You just run the simulation and watch what happens. Sometimes, of course, a simulation will do something that was utterly unexpected. But no matter. If the simulation really was similar to the target system, then the prediction it provides will be a good one. In predictions based on simulations, what you don't know won't hurt you. All of this applies to the off-line simulation theory, of course. If there is some quirk in the human decision-making system, something quite unknown to most people that leads the system to behave in an unexpected way under certain circumstances, the accuracy of predictions based on simulations should not be adversely affected. If you provide the system with the right pretend input, it should simulate (and thus predict) the unexpected output. Adapting a term from Pylyshyn, we might describe this by saying that simulation-based predictions are not 'cognitively penetrable'.¹²

Just the opposite is true for predictions that rely on a theory. If we are making predictions on the basis of a set of laws or principles, and if there are some unexpected aspects of the system's behavior that are not captured by our principles, then our predictions about those aspects of the system's behavior should be less accurate. Theory based predictions are sensitive to what we know and don't know about the laws that govern the system; they *are* cognitively penetrable. This contrast provides a useful way of testing the two theories. If we can find cases in which ignorance about the workings of one's own psychology leads people to make mistakes in predicting what they, or other similarly situated people, would do, it will provide yet another reason to think that the off-line simulation theory is untenable. And, as it happens, cases illustrating cognitive penetrability in the prediction of behavior are not all that hard to find. The literature in cognitive social psychology is full of them. We'll illustrate the point with three examples, but it would be easy to add three dozen more.

First Example: Suppose you are walking through the local shopping mall, and encounter what looks to be yet another consumer product opinion survey. In

this one a polite, well-dressed man invites you to examine an array of familiar products – nightgowns, perhaps, or pantyhose – and to rate their quality. A small reward is offered for your participation – you can keep the garment you select. On examining the products, you find no really significant differences among them. (You couldn't, because, unbeknownst to you they are identical.) What would you do? Confronted with this question, most of us think we would report that the garments looked to be very similar, and then choose one randomly. However, when the experiment was actually tried, this turned out to be mistaken. 'There was a pronounced position effect on evaluations, such that the right-most garments were heavily preferred to the left-most garments.' But it was clear that few of the subjects had any awareness at all of the effect of position on their decision. Indeed, 'when questioned about the effect of the garments' position on their choices, virtually all subjects denied such an influence (usually with a tone of annoyance or of concern for the experimenter's sanity)' (Nisbett and Ross, 1980, p. 207).

This sort of case poses real problems for the off-line simulation theory. Most people have no trouble imagining themselves in the situation described. They can supply their decision-making system with vivid 'pretend' input. But few people who have not heard of the experiment predict that they would behave in the way that the subjects behaved. The natural interpretation of the experiment is that people's predictions about their own behavior (and the subjects' explanations of their own choice) are guided by an incomplete or inaccurate theory, one which includes no information about these so-called 'position effects'.

Second Example: Here's another case to run through your own simulator. Suppose someone in the office is selling \$1.00 tickets for the office lottery. In some cases, when a person agrees to buy a ticket, he or she is simply handed one. In other cases, after agreeing to buy a ticket, the buyer is allowed to choose a ticket from several that the seller has available. On the morning of the lottery, the seller approaches each purchaser and attempts to buy back the ticket. Now imagine yourself in both roles – first as a person who had been handed the ticket, second as a person who had been given a choice. What price would you ask in each case? Would there be any difference between the two cases? On several occasions one of us (Stich) has asked large undergraduate classes to predict what they would do. Almost no one predicts that they would behave the way that people actually do behave. Almost everyone is surprised to hear the actual results.

Ah, yes, the results; we haven't yet told *you* what they are. When the experiment was actually done, 'no-choice subjects sold their tickets back for an average of \$1.96. Choice subjects, who had personally selected their tickets, held out for an average of \$8.67' (Nisbett and Ross, 1980, p. 136). If, like Stich's students, you find this surprising and unexpected, it counts as yet another difficulty for the off-line simulation theory.

Third Example: In the psychology laboratory, and in everyday life, it sometimes happens that people are presented with fairly persuasive evidence that

they have some hitherto unexpected trait. In the light of that evidence people form the belief that they 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 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. If until recently I never had reason to think I had a certain trait, and if the evidence I just acquired has been soundly discredited, then surely it would be silly of me to go away thinking that I *do* have the trait. That seems to be what most people think. And the view was shared by a generation of social psychologists who duped subjects into believing all sorts of things about themselves, 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 is very good at telling real from fake suicide notes, for example, showing her that the evidence was completely phony does not succeed in eliminating the belief. Moreover, third-person observers of the experiment exhibit even stronger 'belief perseverance'. If an observer subject watches a participant subject being duped and then debriefed, the observer, too, will continue to believe that the participant is particularly good at detecting real suicide notes (Nisbett and Ross, 1980, pp. 175–9).

Neither of these results should have been at all surprising to anyone if we predict each other's beliefs and behavior in the way that the off-line simulation theory suggests. But clearly the results *were* both surprising and disturbing. We can't simply ask ourselves what we would do in these circumstances and expect to come up with the right answer. For the theory-theorist, this fact poses no particular problem. When our folk psychology is wrong, it is to be expected that our predictions will be wrong too. It is simply another illustration of cognitive penetrability in predicting and explaining behavior. The theory-theory, unlike the off-line simulation theory, predicts that people's predictions and explanations of behavior will be cognitively penetrable through and through. If it is agreed that these experiments confirm cognitive penetrability, the off-line simulation theory is in serious trouble.

6 Conclusion

Our paper has been long but our conclusion will be brief. The off-line simulation theory poses an intriguing challenge to the dominant paradigm in contemporary cognitive science. Moreover, if it were correct the off-line simulation account of psychological prediction and explanation would largely undermine *both* sides in the eliminativism dispute. But it has been our contention that the prospects for the off-line simulation theory are not very bright. None of the arguments that have been offered in defense of the theory are at all persuasive. And there is lots of experimental evidence that would be

very hard to explain if the off-line simulation account were correct. We don't claim to have provided a knock-down refutation of the off-line simulation theory. Knock-down arguments are hard to come by in cognitive science. But we do claim to have assembled a pretty serious case against the simulation theory. Pending a detailed response, we don't think the off-line simulation theory is one that cognitive scientists or philosophers should take seriously.

Notes

We are grateful to Jerry Fodor for his helpful comments on an earlier version of this paper. Thanks are also due to Joseph Franchi for help in preparing the figures.

- 1 We are grateful to Professor Gordon for providing us with copies of his unpublished papers, and for allowing us to quote from them at some length. Goldmann (1989) is reprinted as ch. 3 in this volume. Gordon (1986) is reprinted as ch. 2; see also ch. 4. Heal (1986) is reprinted as ch. 1.
- 2 The wind tunnel analogy is suggested by Ripstein (1987, p. 475ff). Gordon also mentions the analogy in ch. 4 in this volume, but he puts it to a rather different use.
- 3 The burglar in the basement example is borrowed from Gordon, this volume, p. 62.
- 4 The evidence Gordon cites includes the tendency to mimic other people's facial expressions and overt bodily movements, and the tendency in both humans and other animals to direct one's eyes to the target of a conspecific's gaze.
- 5 Gordon, this volume, pp. 64 ff.
- 6 Ripstein's account of the role of simulation in intentional explanation is quite similar.

I wish to defend the claim that imagining what it would be like to be in 'someone else's shoes' can serve to explain that person's actions... I shall argue that imagining oneself in someone else's situation... allows actions to be explained without recourse to a theory of human behavior. (Ripstein, 1987, p. 465)

[T]he same sort of modeling [that engineers use when they study bridges in wind tunnels] is important to commonsense psychology: I can use my personality to model yours by 'trying on' various combinations of beliefs, desires and character traits. In following an explanation of what you do, I use my personality to determine that the factors mentioned would produce the result in question... I do not need to know how you work because I can rely on the fact that I work in a similar way. My model... underwrites the explanation by demonstrating that particular beliefs and character traits would

lead to particular actions under normal circumstances. (Ripstein, 1987, pp. 476–7)

7 As Jerry Fodor has pointed out to us, the logical geography is actually a bit more complex than figure 5.2 suggests. To see the point, consider the box labeled 'Decision-making (practical reasoning) system' in figure 5.1. Gordon and Goldman tell us relatively little about the contents of this box. They provide no account of how the practical reasoning system goes about the job of producing decisions from beliefs and desires. However, there are some theorists – Fodor assures us that he is one – who believe that the practical reasoning system goes about its business by exploiting an internally represented decision theory. If this is right, then we exploit a tacit theory each time we make a decision based on our beliefs and desires. But now if we make predictions about other people's behavior by taking our own practical reasoning system off-line, then we also exploit a tacit theory when we make these predictions. Thus, contrary to the suggestion in figure 5.2, off-line simulation processes and processes exploiting an internally represented theory are not mutually exclusive, since some off-line simulation processes may also exploit a tacit theory.

In the pages that follow, we propose to be as accommodating as possible to our opponents and to make things as hard as possible for ourselves. It is our contention that prediction, explanation and interpretation of the sorts we have discussed do not use an off-line simulation process, *period*. So if it turns out that Fodor is right (because the practical reasoning system embodies an internally represented theory) *and* that Gordon and Goldman are right (because we predict and explain by taking this system off-line), then we lose, and they win. Also, of course, if Fodor is wrong about how the practical reasoning system works but Gordon and Goldman are right about prediction, explanation and interpretation, again we lose and they win. So as we construe the controversy, it pits those who advocate any version of the off-line simulation account against those who think that prediction, explanation and interpretation are subserved by a tacit theory *stored somewhere other than in the practical reasoning system*. But do keep in mind that we interpret 'theory' broadly. So, for example, if it turns out that there is some non-sentence-like, non-rule-based module which stores the information that is essential to folk-psychological prediction and explanation, and if this module is not used at all in ordinary 'on-line' practical reasoning and decision-making, then we win and they lose.

It might be protested that in drawing the battle lines as we propose to draw them, we are conceding to the opposition a position that they never intended to occupy. As we have already noted, Gordon and Goldman expend a fair amount of effort arguing that a tacit theory is not exploited in folk-psychological prediction and explanation. Since they think that the practical reasoning system is exploited in folk-psychological predic-

tion and explanation, presumably they would deny that the practical reasoning system uses an internally represented decision theory. So it is a bit odd to say that *they* win if Fodor is right about the practical reasoning system and they are right about off-line simulation. This is a point we happily concede. It is a bit odd to draw the battle lines in this way. But in doing so, we are only making things more difficult for ourselves. For we must argue that *however the practical reasoning system works* we do not predict and explain other people's behavior by taking the system off-line. Gordon (unpublished), sec. 3.5.

9 Actually, the developmental facts are rather more complicated than Gordon suggests. For, as Leslie (1988) emphasizes, children are typically able to appreciate and engage in pretend play by the time they are two and a half years old – long before they can handle questions about Maxi and his false beliefs. It is not at all clear how the off-line simulation theory can explain both the early appearance of the ability to pretend and the relatively late appearance of the ability to predict the behavior of people whose beliefs and desires differ from one's own.

10 It's worth noting that both Gordon's account and Leslie's 'predict that only those children who can engage in pretend play can master the concept of belief' (Gordon, this volume, p. 70). This may prove a troublesome implication for both theorists, however. For it is not the case that *all* autistic children do poorly on the Wimmer–Perner test. In the original study reported by Baron–Cohen, Leslie and Frith (1985), only 16 out of 20 autistic subjects failed the Wimmer–Perner test. The other 4 answered correctly. The investigators predicted that these children 'would also show evidence of an ability to pretend play' (p. 43). Unfortunately, no data was reported on the pretend play ability of these subjects. If it should turn out that some autistic children do well on the Wimmer–Perner test *and* lack the ability for pretend play, both Gordon's explanation and Leslie's would be in trouble. If the facts do turn out this way, advocates of the theory-theory will have a variety of other explanations available. But it is much less clear that the off-line simulation account could explain the data, if some autistic children can't pretend but can predict the behavior of people with false beliefs.

11 Another experiment reported by Perner et al. (1987) provides some additional evidence for this conclusion. In the first part of the experiment children were shown a box of Smarties (a type of candy), and asked what they thought was in the box. All of them answered that the box contained Smarties. They were then shown that the box contained a pencil, and no Smarties. After this the children were asked three questions:

- (i) What is in the box?
- (ii) What did you think was in the box when you first saw it?
- (iii) What would a friend, waiting outside, think was in the box if he saw it as it is now?

Most of the younger children answered (iii) incorrectly; they failed to predict their friend's false belief. But more than half of those who got (iii) wrong answered (ii) correctly. They were able to tell the experimenter that they had thought the box contained Smarties, and that they were wrong. In commenting on this experiment, Leslie notes that

[d]espite the ability to *report* their false belief, these 3-year-olds could not understand where that false belief had come from. . . . Despite the fact that they themselves had just undergone the process of getting that false belief, the children were quite unable to understand and reconstruct that process, and thus unable, minutes later, to predict what would happen to their friend. (Leslie, 1988, pp. 33-4)

- 12 Pylyshyn, 1981, 1984. It is perhaps worth noting that we are using the term 'cognitively penetrable' a bit more loosely than Pylyshyn does. But in the present context the difference is not important.

References

- Astington, J., Harris, P. and Olson, D. (eds) 1988: *Developing Theories of Mind*. Cambridge: Cambridge University Press.
- Baron-Cohen, S., Leslie, A. and Frith, U. 1985: Does the autistic child have a 'theory of mind'? *Cognition*, 21, 37-46.
- Chomsky, N. 1965: *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
- Churchland, P. 1981: Eliminative materialism and the propositional attitudes. *Journal of Philosophy*, 78, 67-90.
- Churchland, P. 1989: Folk psychology and the explanation of human behavior. In *A Neurocomputational Perspective*. Cambridge, MA: MIT Press.
- Cummins, R. 1983: *The Nature of Psychological Explanation*. Cambridge, MA: MIT Press.
- D'Andrade, R. 1987: A folk model of the mind. In D. Holland and N. Quinn (eds), *Cultural Models in Language and Thought*. Cambridge: Cambridge University Press.
- Dennett, D. 1969: *Content and Consciousness*. London: Routledge and Kegan Paul.
- Dennett, D. 1978a: Artificial intelligence as philosophy and psychology. In *Brainstorms*. Cambridge, MA: MIT Press.
- Dennett, D. 1978b: Two approaches to mental images. In *Brainstorms*. Cambridge, MA: MIT Press.
- Fodor, J. 1968: The appeal to tacit knowledge in psychological explanation. *Journal of Philosophy*, 65, 627-40.
- Fodor, J. 1975: *The Language of Thought*. New York: Thomas Crowell.
- Fodor, J. 1981: *Representations*. Cambridge, MA: MIT Press.
- Fodor, J., 1987: *Psychosemantics*. Cambridge, MA: MIT Press.
- Fodor, J., Bever, T. and Garrett, M. 1974: *The Psychology of Language: An Introduction to Psycholinguistics and Generative Grammar*. New York: McGraw-Hill.
- Goldman, A. I. 1989: Interpretation psychologized. *Mind and Language*, 4, 161-85. Reprinted as ch. 3 in this volume.
- Gordon, R. M. 1986: Folk psychology as simulation. *Mind and Language*, 1, 158-71. Reprinted as ch. 2 in this volume.
- Gordon, R. M. Unpublished: Fodor's intentional realism and the simulation theory. MS dated 2/90.
- Gregory, R. 1970: *The Intelligent Eye*. New York: McGraw-Hill.
- Greeno, J. 1983: Conceptual entities. In D. Gentner and A. Stevens (eds), *Mental Models*. Hillsdale, NJ: Erlbaum.
- Hayes, P. 1985: The second naive physics manifesto. In J. Hobbs and 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. Reprinted as ch. 1 in this volume.
- Johnson-Laird, P. 1983: *Mental Models: Towards a Cognitive Science of Language, Inference and Consciousness*. Cambridge, MA: Harvard University Press.
- Kahneman, D. and Tversky, A. 1982: The simulation heuristic. In D. Kahneman, P. Slovic and A. Tversky (eds), *Judgment Under Uncertainty*. Cambridge: Cambridge University Press.
- Kosslyn, S. 1981: The medium and the message in mental imagery: A theory. In N. Block (ed.), *Imagery*. Cambridge, MA: MIT Press.
- Leslie, A. 1987: Pretense and representation: The origins of 'theory of mind'. *Psychological Review*, 94, 412-26.
- Leslie, A. 1988: Some implications of pretense for mechanisms underlying the child's theory of mind. In J. Astington, P. Harris and D. Olson (eds), *Developing Theories of Mind*. Cambridge: Cambridge University Press.
- Lycan, W. 1981: Form, function and feel. *Journal of Philosophy*, 78, 24-50.
- Lycan, W. 1988: Toward a homuncular theory of believing. In *Judgement and Justification*. Cambridge: Cambridge University Press.
- Marr, D. 1982: *Vision*. San Francisco: Freeman.
- McCloskey, M. 1983: Naive theories of motion. In D. Gentner and A. Stevens (eds), *Mental Models*. Hillsdale, NJ: Erlbaum.
- Montgomery, R. 1987: Psychologism, folk psychology and one's own case. *Journal for the Theory of Social Behavior*, 17, 195-218.
- Newell, A. 1980: Physical symbol systems. *Cognitive Science*, 4, 135-83.
- Nisbett, R. and Ross, L. 1980: *Human Inference*. Englewood Cliffs, NJ: Prentice-Hall.
- Olson, D., Astington, J. and Harris, P. 1988: Introduction. In J. Astington, P. Harris and D. Olson (eds), *Developing Theories of Mind*. Cambridge: Cambridge University Press.

Perner, J., Leekam, S. and Wimmer, H. 1987: Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology*, 5, 125-37.

Pinker, S. 1989: *Learnability and Cognition*. Cambridge, MA: MIT Press.

Pylyshyn, Z. 1981: The imagery debate: Analog media versus tacit knowledge. In N. Block (ed.), *Imagery*. Cambridge, MA: MIT Press.

Pylyshyn, Z. 1984: *Computation and Cognition*. Cambridge, MA: MIT Press.

Ramsey, W., Stich, S. and Garon, J. 1990: Connectionism, eliminativism and the future of folk psychology. *Philosophical Perspectives*, 4, 499-533.

Ripstein, A. 1987: Explanation and empathy. *Review of Metaphysics*, 40, 465-82.

Rock, I. 1983: *The Logic of Perception*. Cambridge, MA: MIT Press.

Rumelhart, D., Smolensky, P., McClelland, J. and Hinton, G. 1986: Schemata and sequential thought processes in PDP models. In J. McClelland, D.

Rumelhart and the PDP Research Group, *Parallel Distributed Processing*, vol. 2. Cambridge, MA: MIT Press.

Sellars, W. 1963: Empiricism and the philosophy of mind. In *Science, Perception and Reality*. London: Routledge and Kegan Paul.

Stich, S. 1978: Beliefs and subdoxastic states. *Philosophy of Science*, 45, 499-518.

Wimmer, H., Hogrefe, J. and Sodian, B. 1988: A second state in children's conception of mental life: Understanding informational access as origins of knowledge and belief. In J. Astington, P. Harris and D. Olson (eds), *Developing Theories of Mind*. Cambridge: Cambridge University Press.

6

'He Thinks He Knows': And More Developmental Evidence Against the Simulation (Role-taking) Theory

JOSEF PERNER AND DEBORRAH HOWES

1 Introduction

How do children come to understand the mind? There are many different answers in the offing from philosophy of mind. For present purposes we want to focus on the suggestion that children understand the mind by putting themselves in the particular mental state in question by off-line simulation. In developmental circles this is an old idea known under the names of 'role-taking' or 'perspective-taking', which was thought to be the method by which the young child overcomes his egocentric attitude which encourages him to accept his own view as the only one possible (Piaget and Inhelder, 1948/1956, p. 194). This is achieved - as the saying goes - by 'putting himself into the other person's shoes', which, presumably, means experiencing in simulation what the other person experiences for real.

Although simulation is suggested by the terms 'role-' and 'perspective-taking', it has not been worked out as a systematic theoretical position. John Flavell (personal communication), who has used 'role taking' in the title of one of his books (Flavell et al., 1968), assured us that he had not intended to take any particular theoretical position, but used the term purely as the then usual label for the development of social cognition.

Gordon (1986, reprinted as ch. 2 in this volume) and Goldman (1989, reprinted as ch. 3 in this volume) proposed this idea as a systematic position in the philosophy of mind, and Harris (1989, 1991; see also ch. 10 in this volume) turned it into a developmental alternative to the predominant view that children acquire a 'theory of mind'. To see the distinguishing feature of the simulation theory let us consider the following two examples.

Assume you are designing an obstacle course and you are not quite sure whether your wall is the right size. It should be a challenge for climbing it but