which is claimed to advance our understanding still further. But it is, in reality, a step back – either to Cartesianism, in the Goldman-Harris version of it, or to quasi-behaviourism, in the Gordon version. So theory-theory still rules OK!

ACKNOWLEDGEMENTS

I am grateful to the following for their comments on an earlier draft: George Botterill, Jack Copeland, Paul Harris, and Peter J. Smith.

Theories of theories of mind
P. Carruthers + P. Smith (eds)
Cambridge Univ Press (1996)

4 Varieties of off-line simulation

Shaun Nichols, Stephen Stich, Alan Leslie, and David Klein

1 Simulation and information

In the last few years, off-line simulation has become an increasingly important alternative to standard explanations in cognitive science. The contemporary debate began with Gordon (1986) and Goldman's (1989) off-line simulation account of our capacity to predict behaviour. On their view, in predicting people's behaviour we take our own decision-making system 'off line' and supply it with the 'pretend' beliefs and desires of the person whose behaviour we are trying to predict; we then let the decision maker reach a decision on the basis of these pretend inputs. Figure 4.1 offers a 'boxological' version of the off-line simulation theory of behaviour prediction.[1]

The off-line simulation theory of behaviour prediction is a radical departure from the typical explanations of cognitive capacities. In explaining a capacity in some domain (e.g., our ability to solve mathematical problems), the usual strategy in cognitive science is to suppose that the subject has a body of information about that domain (see, e.g., Fodor, 1968a). For example, our ability to predict the motion of projectiles is thought to depend on a body of information about mechanics – a folk physics. Of course, much if not all of these information bases or theories may be tacit or 'sub-doxastic' (Stich, 1978). Further, different theorists have different ideas about how the information is encoded. Some think that the information is encoded in sentence-like structures (Fodor, 1975); others think the information is represented by non-sentential mental models (Johnson-Laird, 1983); still others think that the information is represented by the connection strengths between the nodes of a neural network (Churchland, 1989). Despite the serious and sometimes harsh disagreement between these theorists, they all explain cognitive capacities by appealing to *some* kind of information regarding the domain. In contrast, an off-line simulation account of a capacity claims that the capacity *doesn't* depend on information about the domain. According to such accounts, whether or not a subject has information about the domain is irrelevant to the capacity. Instead of appealing to information about the domain, such theories

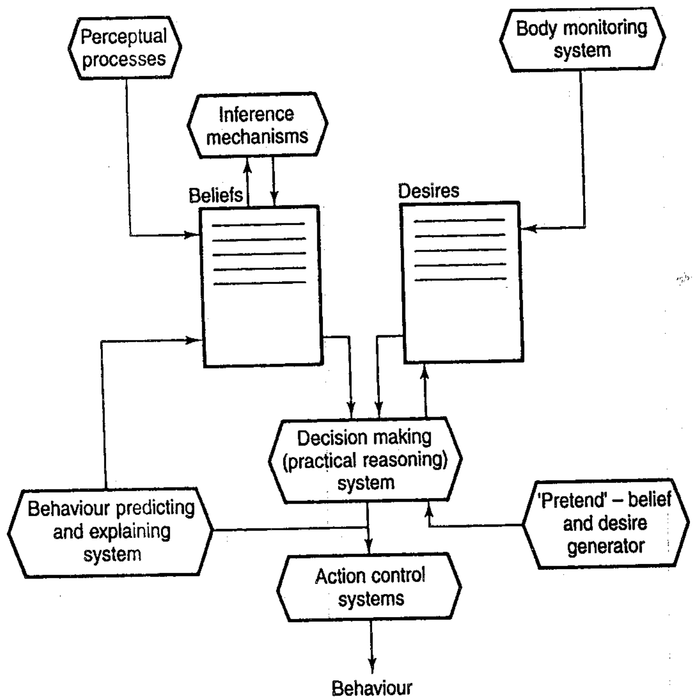

Figure 4.1 Simulation-based account of behaviour prediction

appeal to a mechanism that is already present, but claim that the mechanism is used to support another function. As a result, off-line simulation accounts present a strikingly different picture of cognitive capacities. Indeed, it seems that if off-line simulation can provide plausible accounts of a number of important capacities, it may well constitute a new paradigm in cognitive science.

Given the radical departure off-line simulation takes from information-based theories, simulation may be viewed either as a promising alternative or as an insidious threat to traditional theories in cognitive science. Either way, off-line simulation deserves our careful attention. In this paper, we want to show just how extensive the simulation alternative is. In this first section, we will try to put the notion of off-line simulation into a broader perspective than has been considered in the literature. We will also discuss how cognitive penetrability provides a wedge for empirically determining

whether a capacity requires an information-based account or an off-line simulation account. In the second section, we'll present some empirical results on the cognitive penetrability of our capacity to predict behaviour. In the final section, we'll consider some recent extensions of off-line simulation theory. We'll look at simulation-based accounts of counterfactual reasoning, empathy and mental imagery.

1.1 *Off-line simulation*

Although the off-line simulation theory of behaviour prediction has received the most attention (see, for example, the essays in Davies and Stone, 1995), the basic idea of off-line simulation can be cast in a much broader framework. To explain this framework, it will be helpful to recount a familiar presupposition in cognitive science. One guiding assumption in cognitive science for the past quarter century is that the mind is composed of different cognitive components that are individuated by their functions (Fodor, 1968a, Dennett, 1978). For instance, beliefs are distinguished from desires by the fact that beliefs serve a different function from desires. Hence, cognitive scientists tend to posit distinct belief and desire components. Similarly, we distinguish the capacity to draw theoretical or logical inferences from the capacity to draw practical or desire-based inferences. These capacities too are distinguished by their functions. Distinguishing cognitive components by their functions results in a 'functional architecture' of the mind. A prevalent and colourful way of presenting models of such functional components is by supposing that the mind is made up of various boxes, each of which has a different function.[2] Figure 4.1 is one example of such a 'boxology'.

The basic idea of the off-line simulation theory of behaviour prediction is that the practical reasoning component is taken off line and used for predicting behaviour. However, there's no reason to suppose that the idea of off-line simulation can't be extended to components other than the practical reasoning system. In fact, given a boxology (a functional architecture), *each* component can be viewed as a possible engine of simulation. In principle, *any* component can be taken off line (detached from its usual function) and be used to perform some other function. That is, each component could conceivably be disengaged from its normal purpose and used to support or produce another capacity.

There are a few points worth emphasising about the breadth of this idea. First, it's possible that the same component or mechanism can support several different types of simulation-based capacities. For instance, it's conceivable that the practical reasoning mechanism is taken off line to support our capacity for behaviour prediction as well as our capacity for conditional planning. Indeed, it seems that this is exactly what Goldman has proposed (1993b).

Second, off-line simulation isn't restricted to processing components. The possibility of disengaging a component from its normal functions is clearest for processing or computational mechanisms, like the practical reasoner or the inference device. But it is also true for storage components like the belief box and the desire box.

Previous discussions of off-line simulation accounts have typically characterised the off-line input as 'pretend' input. For example, both Gordon (1986) and Goldman (1989) appeal to pretence in sketching the off-line simulation account of behaviour prediction. Now, it's not entirely clear what 'pretend' is supposed to mean in this context; but perhaps some notion of pretence is essential to characterising an off-line simulation account of behaviour prediction. However, there's no in-principle reason to suppose that *all* off-line simulation theories must be tied to the capacity for pretence. What *is* required is that the component is detached from its normal function. So, the inputs must have a different source than the usual source, but that source needn't be linked to pretence. Indeed, as we've presented it, off-line simulation doesn't necessarily invoke 'inputs' at all. If the mental box that we are taking off line is a processor of some sort (e.g., the practical reasoning device), then inputs are, of course, required. But if the box we are taking off line is a storage system (e.g. the belief box), then it is less than clear that anything appropriately called an 'input' is needed. Consider, for example, using the belief box in the kind of simulation suggested by Paul Harris (1992). Asked to predict what Bill Clinton will think when asked, 'What is the capital of California?', we check to see what we think the answer to the question is. Since we suppose that Clinton is at least as much an expert on United States state capitals as we are, we predict that Clinton will think what we do. There's *some* sort of probe to the belief box here, of course, but it plays a very different role from the pretend inputs posited for practical reasoning or inference simulation.

We've been stressing the potential fertility of off-line simulation. Since virtually any cognitive component is a potential engine of simulation, it seems that off-line simulation accounts might be offered for many different capacities. Indeed, in the last few years, simulation-based accounts *have* been offered for a wide range of cognitive capacities. We have already seen the off-line simulation account of behaviour prediction. On that account, the practical reasoning mechanism is taken off line and used to generate predictions rather than behaviour. We'd now like to sketch two off-line simulation accounts that exploit cognitive mechanisms other than the practical reasoner.

Paul Harris (1992) argues that simulation can provide an excellent account of our capacity to predict the grammaticality judgements of those from the same linguistic community. Suppose you were given a list of grammatical and ungrammatical sentences and asked to predict the judgements other English speakers would make about the grammaticality of the sentences. How do you make such predictions? An information-based account of this capacity would say that you have a body of information about the grammaticality judgements of others in your linguistic community. Harris presents a rather different account of how you would go about predicting such judgements. He writes, '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 judgements for the same reasons' (Harris, 1992, p. 124). According to Harris, then, our capacity to predict grammaticality judgements doesn't depend on our having a body of information about the grammaticality judgements of others. Rather, we use part of our *own* grammaticality system, taken 'off line', to determine the grammaticality judgements of others.

Harris' account of grammaticality judgement predictions suggests an off-line simulation account of a quite different capacity. The prediction of behaviour is the most heralded of our folk psychological capacities, but we also have a capacity to predict what *inferences* others will draw. This capacity would be exploited in the following thought experiment from Stich and Nichols (1995). Suppose you are told the following: 'Sven believes that all Italians like pasta. Sven is introduced to Maria, and he is told that she is Italian.' Now, you are asked to predict what Sven will say if asked whether Maria likes pasta. How do you arrive at your prediction? Well, an information-based account of the capacity would claim that you have a body of information about how people draw inferences. This might be thought of as a tacit theory of reasoning. On the basis of this theory and the information about Sven, you arrive at the conclusion that if asked, 'Does Maria like pasta?', Sven will assent. A quite different account of the capacity is suggested by off-line simulation. On this account, inference prediction proceeds as follows. Hypothetical inputs are fed into your own inference mechanism; the inference mechanism produces the appropriate inference given the pretend inputs; this output is then embedded in the appropriate belief sentence. So, on the above example, you feed your inference mechanism with the pretend beliefs that all Italians like pasta and that Maria is an Italian. Your inference mechanism then produces the conclusion that Maria likes pasta. But this conclusion isn't directly fed into your belief box. Otherwise *you* would come to believe that Maria likes pasta. Rather, this conclusion is embedded in a belief sentence with Sven as the subject. Through this process, you come

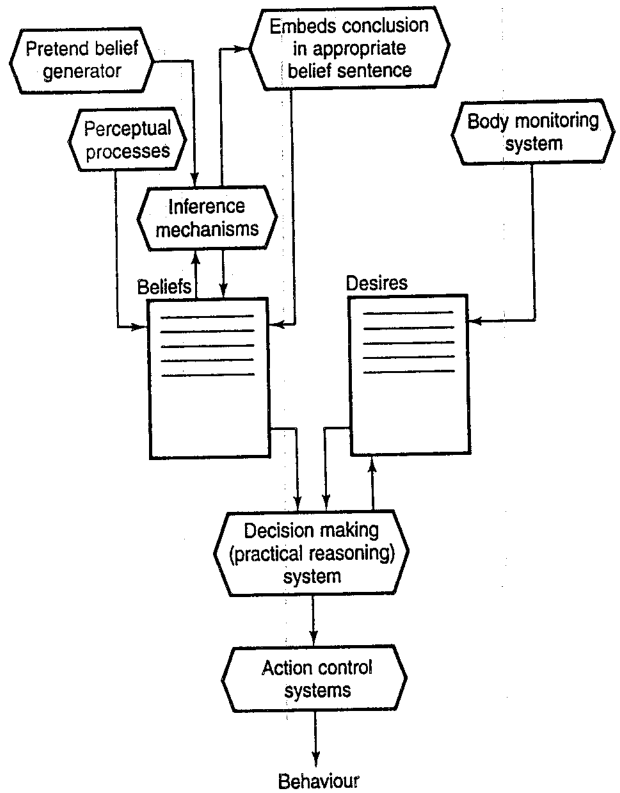

Figure 4.2 Simulation-based account of inference prediction

to believe that Sven will infer that Maria likes pasta. Figure 4.2 is a boxological rendition of this account.

In addition to the foregoing accounts, off-line simulation accounts have been suggested for conditional planning (Goldman, 1992a; Harris, 1992), counterfactual reasoning (Goldman, 1993b), empathy (Goldman, 1993b), and mental imagery (Currie, forthcoming). There is even an account of phoneme recognition from the earliest days of cognitive science that has the relevant features of off-line simulation. Halle and Stevens' analysis by synthesis account of phoneme recognition is perhaps the earliest detailed 'off-line simulation' account in cognitive science (Halle and Stevens, 1962). On

their account, our capacity for phoneme recognition exploits our phoneme *production* system. Crudely put, the idea is that when given phonemic input (i.e., when spoken to), the phoneme production system is taken off line (detached from its normal function) to generate hypotheses for matching the phonemic input. Recognition occurs when a match is found between the phonemic input and one of the hypothetical outputs of the phoneme production system.[3]

So, simulation-based theories can be offered for a wide range of cognitive capacities, and such accounts can provide enticing alternatives to standard information-based accounts. Off-line simulation even makes a sort of evolutionary sense. The literature in evolutionary biology is rife with examples in which a mechanism that was selected for one function, ends up being used for another. For instance, Gould and Vrba report that the wings of the Black Heron serve two important functions. The Heron uses its wings in flight, but it also uses them as a fishing aid. It fishes in shallow water, and it uses its wings to shade the water; this facilitates the bird's ability to see the fish (Gould and Vrba, 1982, pp. 8–9). Clearly, since this can enhance the bird's feeding, it might now be an important selective feature. Gould and Vrba introduce the notion of exaptation, which captures such cases. Exaptations are characters that are 'evolved for other usages . . . and later "coopted" for their current role' (Gould and Vrba, 1982, p. 6). Or, as Vrba puts it, 'An exaptation is a character that is currently useful (and subject to selection) in a particular role, but that was not shaped by step-by-step selection for that role' (Vrba, 1989, p. 130). Gould and Vrba argue that exaptations form an extremely important category in evolutionary biology. If we import the notion of exaptation into evolutionary psychology, then capacities produced by off-line simulation look to be excellent candidates for exaptations. For instance, in the off-line account of behaviour prediction, a psychological mechanism that was built (selected) for one purpose is coopted for another.

Off-line simulation can offer accounts for a wide array of cognitive capacities. And it's quite possible that in some cases the simulation story is appropriate while in other cases it's misguided. Indeed, we suspect that some cognitive capacities probably *are* subserved by off-line simulation. For instance, Harris' simulation account of the capacity to predict grammaticality judgements is quite plausible. However, we are still extremely sceptical of the off-line simulation theory of behaviour prediction, and in section 2 we will present some empirical results against the theory. But before we present our data, we need to get clear about how to empirically distinguish off-line simulation accounts from more traditional information-based accounts.

1.2 *Information, simulation, and cognitive penetrability*

Stich and Nichols (1992) argued that demonstrating that behaviour prediction is 'cognitively penetrable' would be strong evidence that behaviour prediction derives from a theory or information base rather than off-line simulation. A capacity is cognitively penetrable in this sense if that capacity is affected by the subject's knowledge or ignorance of the domain. The point was that if behaviour prediction derives from off-line simulation, then the subject's knowledge or ignorance of generalisations about human behaviour should be irrelevant to the subject's performance on behaviour prediction tasks. However, the relevance of cognitive penetrability for the simulation debate extends far beyond the issue over behaviour prediction. We would maintain that, for any cognitive capacity, demonstrating that that capacity is cognitively penetrable indicates that the capacity derives from an information base rather than from off-line simulation.

For present purposes, the crucial difference between information-based accounts and simulation-based accounts is that an information-based account claims that the capacity depends on information about the domain but a simulation-based account claims that information about the domain is irrelevant to the capacity. This difference is exactly what matters for cognitive penetrability. Insofar as a capacity is affected by the subject's knowledge or ignorance of the domain, that capacity is 'cognitively penetrable'. As a result, any capacity that depends on information about the domain will be cognitively penetrable. By contrast, a simulation-based account of a capacity claims that the subject's information about the domain is irrelevant. Accordingly, if a capacity derives from off-line simulation, that capacity should not be cognitively penetrable. In other words, the subject's knowledge or ignorance of the domain should be irrelevant to their performance on tasks exploiting the capacity.

To avoid confusion over the matter, we want to clarify a difference between our notion of cognitive penetrability and Pylyshyn's notion (1980). On Pylyshyn's notion, a capacity is cognitively impenetrable insofar as it is unaffected by the subject's beliefs and desires outside of the information base devoted to the domain. The notion of cognitive penetrability important to the simulation debate is rather that a capacity is cognitively impenetrable only if the subject's cumulative knowledge or ignorance of the domain is irrelevant to the subject's performance on tasks exploiting the capacity. It's perfectly conceivable, then, that a capacity might be cognitively impenetrable in Pylyshyn's sense without being cognitively impenetrable in the sense relevant to simulation. For example, it's possible that our capacity to predict behaviour depends on a folk psychology module in Fodor's sense (1983). If, so then folk psychology would be both a theory

(and thus cognitively penetrable in our sense) and cognitively impenetrable in Pylyshyn's sense.

Perhaps the easiest way to test whether a capacity is cognitively penetrable in our sense is to test subjects on a task for which they plausibly lack the relevant information about the domain. According to an information-based account, where subjects lack important information about the domain, they will perform poorly. On the other hand, according to simulation-based accounts, since subjects don't depend on information about the domain to perform the tasks, the subject's ignorance about the domain should make no difference to the subject's performance.

We've been urging that off-line simulation offers an alternative to standard information-based explanations. Furthermore, for a wide range of capacities including behaviour prediction and inference prediction, these two possibilities exhaust the current playing field. As a result, in many cases, cognitive penetrability seems to be a decisive test. Showing that a capacity is cognitively penetrable suffices to show that that capacity requires an information-based account. For to show that a capacity is cognitively penetrable is just to show that the capacity depends on information about the domain. However, if a capacity is not cognitively penetrable, then an information-based account is inappropriate, and typically the only other option is a simulation-based theory.

Off-line simulation is not, of course, the only kind of simulation. Another familiar notion of simulation comes from computer science. On this notion of simulation, one tries to predict the behaviour of a system by exploiting a computer model of the system (e.g., Widman and Loparo, 1989, p. 15). Such models may be characterised by mathematical equations or rules. For example, suppose we wanted to be able to determine how long it would take a forest fire in Yellowstone to reach surrounding towns under prevailing conditions. The speed at which the fire will burn depends partly on the moisture conditions, the wind conditions, and the density of the forest. As a result, a good computer model would include equations that capture the impact of these variable conditions as well as information about the size of the forest and the location of the towns. This notion of simulation relies on a model that is constituted by an internally represented body of information; as a result, we'll call it information-based simulation.

Information-based simulation differs from off-line simulation in important ways. As we have taken pains to point out, one way to demonstrate that a capacity derives from off-line simulation is to show that the capacity does not depend on a body of information about the domain. However, to show that a capacity depends on information-based simulation, one would have to show that the subject *is* exploiting a body of information about the domain. Simulation theorists sometimes seem to conflate the distinction. For example, as Stich and Nichols (1992) note, Goldman enlists research on information-based simulation as support for off–line simulation. Goldman writes:

> several cognitive scientists have recently endorsed the idea of mental simulation as one cognitive heuristic, although these researchers stress its use for knowledge in general, not specifically knowledge of others' mental states. Kahneman and Tversky, 1982, propose that people often try to answer questions about the world by an operation that resembles the running of a simulation model. The starting conditions for a 'run', they say, can either be left at realistic default values or modified to assume some special contingency (Goldman, 1989, p. 174).

Gregory Currie also invokes information-based simulation in the debate over off-line simulation. He writes: 'What goes on in the simulator is a substitute for real action; it gives us, under optimal conditions, the information that action would give us about the success or failure of a strategy, without the costs of failure' (Currie, 1995, p. 10). Currie defends his appeal to information-based simulation as follows: 'Simulation theorists have stressed that simulation may employ theory without ceasing to be simulation (see e.g. Goldman's distinction between 'process' and 'theory' driven simulation in his 1992). But this point continues to be overlooked by opponents of simulation theory' (Currie, 1995, p. 13, n. 16).

We would scarcely deny that there are important similarities between the two kinds of simulation. However, in the present context, it's extremely important to keep the distinction clear and prominent. In the first place, as we've noted, information-based simulation theories have radically different empirical commitments than off-line simulation theories. Information-based theories *must* exploit information about the domain. Off-line theories on the other hand, expressly do not exploit information about the domain. So, in order to evaluate a simulation theory, it's vital that we know whether the proposal is information-based or off-line.

Further, it's relatively uncontroversial that we have the capacity for information-based mental simulation. There are, of course, a number of unresolved questions about this kind of mental simulation. There are debates about how the models are encoded and what mechanisms underlie the capacity. However, virtually no one denies that we *have* such a capacity. By contrast, the claim that we have a capacity to take our decision making mechanism off line is enormously controversial.

In this section, we've been concerned to clarify the theoretical groundwork for off-line simulation. In the next section, we take an empirical turn. We will present results on the cognitive penetrability of behaviour prediction.

2 The cognitive penetrability of behaviour prediction: some experimental results

When subjects do a good job at predicting the behaviour of other people, it is often difficult to determine whether they are relying on simulation or internalised information. For if they are relying on internalised information, it may well be the case that this information is not readily available to conscious access. Thus the mere fact that subjects can't report the relevant information, or recognise various characterisations of the information they are using, provides no reason to conclude that they are not invoking a tacit theory. When subjects do a poor job of predicting the behaviour of other people, however, the situation is quite different. For in these cases, the explanatory resources of the theory-theory are considerably greater than those of the simulation theory. On a simulation account like the one portrayed in Figure 1, mistaken predictions can arise in one of two ways.

(i) The predictor's Decision Making (or Practical Reasoning) mechanism is different from the target's.
(ii) The Pretend Belief and Desire Generator has not provided the Decision Making System with the right pretend beliefs and desires – i.e. with the ones that actually motivate the target person whose behaviour is to be predicted.

If an experimental situation can be designed in which subjects systematically mispredict the behaviour of targets, and in which it is unlikely that either (i) or (ii) is the source of the problem, then the simulation account will be seriously challenged. Often, however, these cases will be easily accommodated by the theory-theory, which can attribute the error to the fact that the internalised information on which the subject relies is mistaken or incomplete. If there is some psychological process that is unknown to folk psychology, and if this process affects people's behaviour in a given sort of situation, then it is not surprising that subjects who rely on folk psychology to predict how others will behave in that situation will be mistaken in their predictions.

Stich and Nichols (1992) reported an informal experiment that turned on just such an unsuspected psychological phenomenon. The phenomenon in question was first reported by Ellen Langer (1975). She called it 'the illusion of control', but we prefer a less theory-laden label; we'll call it 'the Langer Effect'. In one of her experiments, Langer organised a football pool in the office of an insurance agency a few days prior to the Superbowl, selling tickets at $1.00 each. Some subjects were offered a choice of tickets; others were offered only one. The day before the big game, Langer said she would be willing to buy the tickets back from the subjects, and asked how much they wanted for them. The surprising result was that subjects who had no choice of tickets sold them back for an average price of $1.96, while subjects who had a choice sold theirs back for an average price of $8.67.

In their informal experiment, Stich and Nichols (1992) read a description of Langer's experiment to a group of undergraduates and asked them to predict what Langer's subjects would do. Not surprisingly, the students got it wrong. They predicted no significant difference between the price asked by the no-choice subjects and the price asked by the subjects who were given a choice. Stich and Nichols tried to use this result as an argument against simulation accounts of behaviour prediction. But the Simulation Theorists were not convinced. Both Harris (1992) and Goldman (1992b) complained that the way in which the facts were presented to the students made it very unlikely that they would use the right pretend beliefs and desires in generating their predictions. Each participant in Langer's experiment was exposed to only one condition – either choice or no-choice. And there was a delay of several days between buying the ticket and being asked to sell it back. In contrast, the students who were asked to predict how Langer's subjects would behave were told about both conditions, and the time delay between being told about the purchase and being asked to predict the sell-back price was only a minute or two. Given these differences, Harris and Goldman protested, it would hardly be surprising if the students used the wrong pretend-inputs in making their prediction. In a later paper Stich and Nichols (1995) concede that the criticism is a fair one. The experiments we are about to report were designed to side-step the shortcomings in Stich and Nichols' informal experiment.

The first issue that needs to be addressed is the importance of the time lag in Langer's original experiment. Was this an essential factor in producing Langer's effect? To answer this question, we conducted an experiment similar to Langer's, but with no significant time lag. Subjects, who met one-on-one with the experimenter, were told that they would be asked to judge the grammaticality or ungrammaticality of fifteen sentences to be read by the experimenter. Before reading them the sentences, it was explained that to reward them for their participation, a lottery had been arranged. The prize was $30.00. At this point, some of the subjects were given a numbered lottery ticket by the experimenter. Other subjects were invited to select one of three lottery tickets that were offered. The experimenter then read the fifteen sentences, and had subjects record their judgements on an answer sheet. The grammar-judgement task lasted about five minutes. When it was complete, the subjects were told that it might be necessary to run more subjects than planned. And since the experimenter wanted to give all subjects a reasonable chance at winning the lottery, he might want to buy back some of the lottery tickets. Subjects were asked to set a price at which they would be prepared to sell the lottery ticket back to the experimenter, and to record that price on their answer sheet. There were a total of thirty subjects, fifteen in each condition. The mean price set by subjects in the no-choice condition was $1.60; the mean price set by subjects in the choice condition was $6.29. ($p < .05$). These results clearly indicate that the Langer Effect can be obtained without any significant delay between the time the subjects receive the lottery tickets and the time they are asked for a price at which they will sell them back.

The next question to be addressed is whether observers can accurately simulate the decision making process that leads subjects in Langer-style experiments to set the prices they do. In order to assist observer-subjects in generating the best possible pretend inputs (if that is indeed what they do) we produced a pair of video tape recordings of a subject (actually a confederate) participating in the experimental procedures just described. The two tapes were identical except for the first two minutes in which the grammatical judgement task was explained and the lottery ticket was given (on one tape) or chosen (on the other tape). Each tape was shown to a separate group of observer-subjects. The observer-subjects were provided with answer sheets identical to the one provided to the subjects in the Langer-style experiment. Observer-subjects were asked to predict what the experimental subject on the tape would say about the grammaticality or ungrammaticality of each of the fifteen sentences. They were also asked to predict the price the subject on the tape would set for selling back his lottery ticket. There were 38 observer-subjects in the choice condition and 39 in the no-choice condition. The mean predicted buy back price was $7.82 in the choice condition and $9.37 in the no-choice condition. (The difference is not statistically significant.) These numbers are quite high, since several less than co-operative observer-subjects in each group gave absurdly high numbers. To correct for this, we reanalysed the data after discarding all predictions higher than $15.00. On this analysis, the mean in the choice condition was $4.62; the mean in the no-choice was $3.47. (Once again, the difference is not statistically significant.) These results are summarised in Table 4.1. For completeness we also calculated the success rate of observer-subjects on the grammatical judgement prediction task. All of the authors judged all of the sentences used to be clearly grammatical or clearly ungrammatical. Using our judgements as a criterion of correctness, the observers predicted correctly 84% of the time, with no significant variation between the two conditions.

It is our contention that these experimental results pose a serious problem for those who think that behaviour prediction relies on the sort of simulation sketched in figure 1. For, as we noted earlier, this sort of simulation has only two straightforward ways of accounting for systematically mistaken predictions. Either the predictors' decision-making system differs from the targets', or the predictors are providing their decision-making

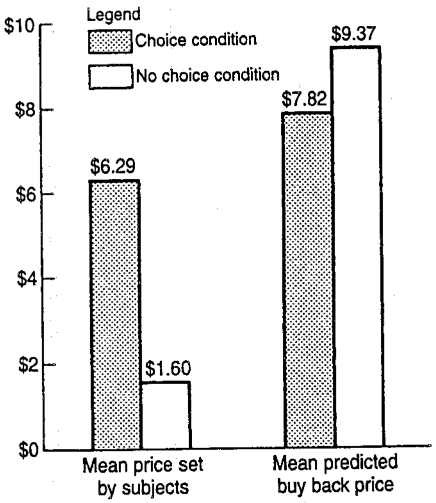

Table 4.1

system with the wrong pretend inputs. In the present experiment, the first option is very implausible, since the predictors and the subjects in the Langer-style experiment were drawn from the same population. There is every reason to believe that if the predictors had been subjects in the Langer-style experiment they would have behaved in much the same way that the actual subjects did. Moreover, the conditions for producing pretend inputs were about as good as they could possibly be, since predictors observed the target on video tape as he was making his decision. Moreover, post-experiment interviews with a number of subjects indicated that all of them correctly remembered whether or not the target they observed had been given a choice of lottery tickets. On the theory-theory account of how behavioural predictions work, these results are easy enough to explain. Folk psychology includes no information about the Langer Effect, so predictors get it wrong. Perhaps Simulation Theorists can produce a plausible alternative explanation of these results. But we haven't a clue what it might be.

3 Off-line simulation: new directions

The results reported in the last section make us sceptical of the off-line simulation account of behaviour prediction. However, off-line simulation

might provide a plausible account of other important capacities. In the first section we stressed the fertility of the idea of off-line simulation. And in the last few years, off-line simulation accounts have been offered for three crucial but strikingly different capacities: counterfactual reasoning, empathy, and mental imagery. In this section, we want to discuss each of these capacities in relation to the simulation debate. We don't, however, propose to decide whether off-line simulation is the correct account for any of these capacities. We only hope to chart the philosophical and psychological space.

3.1 Counterfactual reasoning

Counterfactuals play an enormous and vital role in our inferential lives. Counterfactual thoughts range from the pedestrian – if I'd brought my umbrella I wouldn't be wet – to the profound – if I travelled at the speed of light, I'd turn into energy. The importance of counterfactuals is reflected in the amount of attention they've received in both philosophy and psychology. In the philosophy of language, there is an extensive literature on counterfactuals (e.g., Goodman, 1947; Stalnaker, 1968; Lewis, 1973). There is also a sizeable literature in cognitive social psychology on the generation of counterfactuals (e.g., Kahneman and Tversky, 1982; Markman *et al.*, 1993; Roese and Olson, 1993). In the philosophy of language, there are two closely related projects concerning counterfactuals. One project is to provide an account of the truth conditions for counterfactual conditionals. The goal is to determine what makes such sentences true or false. This project parallels the attempt to provide a semantics for belief sentences. In addition to outlining the semantics of counterfactual conditionals, philosophers of language have tried to produce a logic for counterfactuals. That is, philosophers have tried to determine which inferences are valid given counterfactual premises. This project parallels the attempt to provide a logic for material conditionals, modal operators, temporal terms, etc. For each of these traditional philosophical endeavours, there is a complementary psychological project. Whereas philosophers of language have discussed how counterfactual conditionals *should* be evaluated, the psychological issue is how people *actually* go about evaluating counterfactuals. Similarly, while logicians try to determine what inferences are valid given counterfactual premises, the psychological question is what kinds of inferences people are willing and unwilling to make given counterfactual premises.[4]

Simulation theory offers an extremely interesting perspective on the psychological projects. The question posed by the simulation debate concerns the psychological mechanisms responsible for counterfactual reasoning.

When people actually evaluate counterfactuals or make inferences over counterfactuals, are they using some kind of off-line simulation, or are they guided by a body of counterfactual-specific rules or information?

Goldman (1992a) assumes that counterfactual reasoning requires an off-line simulation account. By 'counterfactual reasoning', he seems to mean the evaluation of counterfactual conditionals. He writes, 'When considering the truth value of "If X were the case, then Y would obtain", a reasoner feigns a belief in X and reasons about Y under that pretence' (1992a, p. 24). Goldman likens this kind of counterfactual reasoning to off-line simulation of behaviour prediction. He takes counterfactual reasoning to be an example of a process that exploits the same mechanisms as the off-line account of behaviour prediction (1992a, pp. 22–4).

We are sceptical of Goldman's view that the evaluation of counterfactual conditionals can be assimilated to the off-line simulation of behaviour prediction. As mentioned in section 1, it's common to distinguish the mechanisms devoted to theoretical reasoning from those devoted to decision making. In figure 4.1, this distinction is depicted by separate boxes for the decision making system and the inference mechanisms. The off-line simulation account of behaviour prediction depends on the decision making system, not the inference mechanisms. However, evaluating counterfactuals seems to be a process by which we come to beliefs, not a process by which we come to decisions. That is, the evaluation of counterfactuals, unlike off-line simulation of behaviour prediction, concerns theoretical inference mechanisms, not the decision-making system.

Although off-line simulation of behaviour prediction provides a poor model for the evaluation of counterfactuals, off-line simulation of *inference* prediction does provide a credible model for how we evaluate counterfactuals. If we suppose that the reasoning processes that are taken off line are the theoretical reasoning processes, then Goldman's account becomes much more plausible. On such an account, we evaluate sentences of the form 'If X were the case, then Y would obtain' as follows: Our inferential mechanisms are taken off line and fed the pretend belief that X. Given that pretend belief, we let our inferential mechanisms work off line to see what can be concluded about Y. In keeping with the off-line simulation theory, we suppose that the inferential mechanisms operate over that feigned belief as if it were a genuine belief.

We think that off-line simulation provides a plausible and interesting account of how we evaluate counterfactual conditionals. But it doesn't have a corner on the market – there are other plausible models of this capacity that don't derive from off-line simulation. One possibility is that the evaluation of counterfactuals depends on a body of meta-linguistic principles concerning entailment relations among sentences. Such principles would be

stored in the belief box; however, like the principles of folk physics and folk psychology, they may not be readily accessible to consciousness. According to this meta-linguistic theory, we know, perhaps tacitly, that certain sentences follow from other sentences. And we can see, as a result, that if some sentences were true, other sentences would also be true. To clarify the nature of this account, let's consider an example. Most would take the following counterfactual to be true:

> If Bill Clinton had lost the election, Hillary Clinton would not be the First Lady.

On the current proposal, we evaluate this sentence in the following way. We know that the following sentences are true: 'The First Lady is the wife of the President'; 'Hillary Clinton is Bill Clinton's wife'; 'The loser of the election doesn't become President.' In addition, we know various entailment relations among sentences. Given our knowledge of the entailment relations, plus our knowledge of the true sentences, we reason that if 'Bill Clinton lost the election' were true, 'Hillary Clinton is not the First Lady' would also be true. This account involves neither pretence nor taking the inference mechanisms off line. Rather, it relies on a body of information about relations between sentences.

Another possible alternative to the simulation account is that the evaluation of counterfactual conditionals depends on a body of counterfactual-specific rules in the inference mechanism. Just as we have a set of rules that guides our evaluation of the material conditional, so too, according to this view, we have a set of inference rules devoted to evaluating counterfactual conditionals. Again, such an account needn't appeal to pretence nor to taking the inference mechanisms off line.

It is, of course, an interesting question which, if any, of these accounts is right. But it's not our purpose here to decide that. Indeed, we suspect that given the current state of the evidence, it's difficult to determine with any confidence whether off-line simulation provides an adequate account of how we evaluate counterfactuals. One way to test the off-line simulation account under consideration might be by exploring how well autistic people are able to evaluate counterfactuals. We have followed Goldman in supposing that the off-line input would be 'pretend' input. One of the central features of autism is the lack of pretend play, and both Goldman (1989) and Gordon (1986) argue that the fact that autistic people fail theory of mind tasks indicates that pretence is involved in those tasks. We're sceptical that the performance of autistic people on theory of mind tasks can be fully explained by their problems with pretence. However, we do think that studying autistics' reasoning about counterfactuals might provide some evidence to help assess Goldman's off-line account of how we evaluate counterfactual conditionals. If the capacity to evaluate counterfactuals depends on the capacity for pretence as Goldman suggests, then autistic people should perform poorly on tasks that require them to evaluate counterfactuals. One way to run such an experiment would be to assemble a list of counterfactual conditionals and see whether autistic people do as well as their mental peers at determining whether the sentences are true or false. According to the off-line account under consideration, autistic people should perform poorly on this task.

Leslie (1987b, 1988a, 1994a) has detailed a quite different set of connections between pretence, autism, and theory of mind. Leslie shows that the capacity for pretence emerges very early in development between 18 and 24 months of age. He points to a critical feature of this capacity, namely, the 'yoking' in development between the capacity for pretending by oneself and understanding pretence-in-others. As soon as the child is able deliberately to entertain counterfactual suppositions as evidenced by solitary pretence, she is also able to understand when other people deliberately entertain counterfactual suppositions, that is, pretend. By two years of age, an infant can share pretend scenarios with others and demonstrate that she understands the specific content of the other person's pretence. In this regard, pretending is strikingly different from believing. Infants only a few months old have beliefs (about the mechanical properties of hidden objects, for example) but, as far as anyone can tell, they are not able to understand beliefs-in-others until a considerable time later – about two years later on the most optimistic estimate.

Leslie's model accounts for the above facts in the following way. In terms of 'boxology', having a belief can be thought of as placing a representation in the 'belief box' – that is, as a particular kind of functional relation between the organism and one of its representations; while having a desire is placing a representation in the 'desire box' – that is, as another kind of functional relation between the organism and its representations. Since young infants, by assumption, can both have beliefs and have desires, we assume that early on infants possess a belief box, a desire box and some representations. But, according to Leslie, in terms of boxology, there is no such thing as the 'pretend box', and thus no such thing as simply 'having a pretend'. Instead, pretending is a special case of placing a representation in the 'belief box', where the representation says in effect, 'someone is pretending such and such'. In a system of this type, solitary pretence and understanding pretence-in-others are inevitably yoked, whereas having beliefs and understanding beliefs-in-others are not. Like pretence, understanding beliefs-in-others depends upon the development of representations of propositional attitudes (what Leslie calls the 'metarepresentation').

Leslie's metarepresentational model has had an important application to understanding abnormal development. Autistic children show a set of behavioural abnormalities that can occur even if the child has a borderline to normal IQ level. These behavioural signs (Wing and Gould's (1979) 'triad') comprise social incompetence, communicative impairment and a lack of normal pretend play. Putting these signs together with the above model led to the hypothesis that autistic children are metarepresentationally impaired. This in turn led to the prediction that autistic children would show a specific impairment in their understanding of belief. This impairment would be specific in at least two senses: first, children with autism would be impaired relative to their own general intellectual level; and second, other groups of children with comparable or lower IQ levels would not be similarly impaired. The first experimental studies of this conjecture supported the prediction (Baron-Cohen, *et al.*, 1985, 1986) showing that only 20% of a group of autistic children with a mean IQ of 82 passed a test of false-belief understanding whereas 86% of a group of Down's syndrome children with mean IQ of 64 passed. These results have subsequently been confirmed and extended by numerous studies around the world (e.g., Baron-Cohen, 1989a, 1991a; Leslie and Frith, 1988; Leslie and Thaiss, 1992; Ozonoff *et al.*, 1991a; Perner *et al.*, 1989; Reed and Peterson, 1990; Roth and Leslie, 1991; Sodian and Frith, 1992; for a short review see Leslie, 1992; and for a discussion of some current issues see Leslie and Roth, 1993). Of particular relevance to the present discussion are two recent studies which we shall describe. Leslie and Thaiss (1992) controlled for the possibility that autistic children might fail tests of false belief understanding because they are unable to meet one or some of the general problem solving or general processing demands made by such tasks. A standard task (e.g. Baron-Cohen *et al.*, 1985) has Sally place a marble in a basket and go away for a walk. While she is away, Ann removes the marble and places it in a box. The child is asked about where Sally put the marble in the beginning, where it is now, and finally where Sally will look for the marble on her return. It makes no difference to the results if children are asked where Sally thinks the marble is rather than where she will look. Such a task no doubt makes numerous general demands upon the solver. Working memory is required, language processing skills, abstract reasoning, perhaps mental imagery, 'executive functions' of various kinds and perhaps other, as yet unthought of, processes are required. How can we address so many possibilities and support the idea of an impairment specific to understanding mental states?

Fortunately there is a way. Zaitchik (1990) reported an elegant test of understanding out-of-date photographs that almost exactly mirrors the general problem solving structure of false-belief tasks. In Zaitchik's task, Sally is replaced with a Polaroid camera and Sally's belief is replaced by a Polaroid photograph. The marble is placed in the basket as before, but, instead of Sally forming a belief about the marble in the basket, the camera forms a photograph of the marble in the basket. The photograph is then placed face down on the table before the child can see it (after all the child does not get to see Sally's belief!). The marble is then removed from the basket and placed in the box. Now the child is asked where the marble was when the photograph was taken, where the marble is now really, and finally, where, in the photograph, is the marble? For normal children, these two tasks, out-of-date photographs and out-of-date beliefs are nearly equivalently hard. Most normal three-year-olds fail both tasks, while most normal four-year-olds pass both tasks. However, experimental analysis shows that for normal children there is a small but reliable effect that if they pass only one of these tasks, it is false belief rather than photographs. Autistic children, by contrast, while failing false belief, perform at or near ceiling on photographs tasks. Similar results are obtained if simple maps or drawings are used instead of a camera and photographs (Charman and Baron-Cohen, 1992; Leslie and Thaiss, 1992). If false-belief tasks require 'imagery' for their solution, then presumably so do the photographs, maps, and drawings tasks. Of course, 'simulation', specifically of mental states, might be the problem. But, as Leslie and German (1994) show, the best explanation for that remains impaired metarepresentational processing.

Although the Leslie and Thaiss (1992) results above undermine the idea that autistic children fail false-belief tasks because they cannot process counterfactual information, a recent study has looked at this directly. Scott *et al.* (1994) tested normal four-year-olds, Down's syndrome children and children with autism on a task that required them to solve simple counterfactual syllogisms. For example, the children were told 'All pigs can fly. Porky is a pig' and then asked 'Can Porky fly?' The normal four-year-olds and Down's syndrome children performed rather poorly on such syllogisms showing a lot of 'reality intrusion' errors. The autistic children, on the other hand, performed very well, clearly demonstrating their mental age advantage. In an attempt to help the normal and Down's children, the children were asked to form a picture of a pig in their heads and then to make this 'pig in the head' fly. To determine what the children had understood about this talk of 'pictures in the head', they were asked whether the pig was real or 'just in their head' and whether the experimenter could see the pig in their head. Most of the normal and Down's children correctly answered that the pig was 'just in their head', and, no, the experimenter could not see it. However, less than half the autistic children got these questions right. They seemed to be confused by this discourse concerning a mental entity. The children were then presented with another set of counterfactual syllogisms, but this time they were asked to form a picture in the head of the situation described. Now the normal and Down's children appeared to be released from their previous reality-bound responses, apparently appreciating that the discourse was about counterfactuals. For young normal children, thinking about mental states facilitates access to their counterfactual reasoning abilities. The autistic children, by contrast, performed drastically worse under these conditions, apparently becoming confused and showing reality intrusion errors for the first time. Counterfactual reasoning is well within the capabilities of autistic children tested; nevertheless, they had difficulty following discourse about mental states. Their difficulty with this task cannot be attributed to difficulty forming and transforming mental imagery given the results of Leslie and Thaiss (1992) and, more directly, given their normal performance on standard mental rotation tasks (Shah, 1988). As Leslie and Roth (1993) point out, metarepresentational impairment succinctly accounts for both the specific disabilities and the spared abilities in the syndrome of childhood autism. The same model also provides the only adequate account of early pretence, linking this to the capacity to acquire a theory of mind.

3.2 *Empathy*

As counterfactual reasoning is vital to our inferential lives, so empathy is an integral part of our emotional lives. Further, according to many writers, empathy is essential for moral reasoning and certain types of aesthetic experience. Although there is no widely accepted definition of empathy, most researchers agree that, at a minimum, empathy is a 'vicarious sharing of affect' or 'an emotional response that stems from another's emotional state or condition and that is congruent with the other's emotional state or situation' (Eisenberg and Strayer, 1987, pp. 3,5). The literature on empathy takes empathy to cover a broad range of behaviour, from reactive new-born cry and motor mimicry to deliberate role taking (Hoffman, 1984, 1987). Goldman (1993b) maintains that empathic phenomena demand a simulation account. In this subsection, we want to explore the extent to which empathy is a product of simulation. After presenting simulation-based and information-based approaches to empathy, we'll consider how well these approaches can accommodate the empathic phenomena.

3.2.1 Two models of empathy

There are, of course, many distinctions to draw among different accounts of empathy. But for present purposes, the distinction we're concerned with is between simulation-based theories and information-based theories. Before we draw this distinction, though, we should give some indication of how non-empathic emotion is supposed to work.

There is a spirited debate in psychology over the causal basis of emotions. The debate revolves around the relative contributions of physiological and cognitive factors. Still, it's fairly uncontroversial that in many instances, emotions are at least partially the product of beliefs or memories. For example, a person will likely experience fear on discovering that he is being pursued by a bear. Or, a person might feel sorrow on reminiscing about a deceased pet. In such cases, it's likely that an emotional response system takes beliefs or memories as inputs, and produces emotional states as outputs. This understanding of emotion seems to be presumed by Goldman's simulation account of empathy. In our discussion, we too will assume this conception of emotion.

Empathy as off-line simulation. Goldman places empathy in the context of simulation as follows:

> The simulation process consists first of taking the perspective of another person, i.e., imaginatively assuming one or more of the other person's mental states. Such perspective taking might be instigated by observing that person's situation and behaviour or by simply being told about them, as when one reads a history book or a novel. Psychological processes then (automatically) operate on the initial 'pretend' states to generate further states that (in favourable cases) are similar to, or homologous to, the target person's states. I conceive of empathy as a special case of the simulation process in which the output states are *affective* or *emotional* states rather than purely cognitive or conative states like believing or desiring (1993b, p. 141)

Goldman's remarks suggest the following simulation account of empathy. A pretend-belief generator feeds into our emotional response system. The emotional response system operates on that feigned belief just as it would operate on an unfeigned belief. The output from the emotional response system is then the same emotional response as it would have been if the belief weren't feigned (see fig 4.3).

This simulation account of empathy is, of course, quite different from the simulation account of behaviour prediction. In the first place, the mechanism used to support empathy is presumably the emotional response system, rather than the practical reasoning system. When empathising, we don't reason to the conclusion that we should feel a certain way. We don't *decide* how to feel. Further, unlike the off-line simulation of behaviour prediction, on the present theory, the output from the mechanism is *not* taken off line. The output from the emotional response system is a genuine emotional response. This account, then, is not 'off-line' at both ends. Rather, we might say that it's 'input deviant'. This makes the account a special case of simulation, as Goldman says. Nonetheless, it's an important application of the basic idea behind off-line simulation.

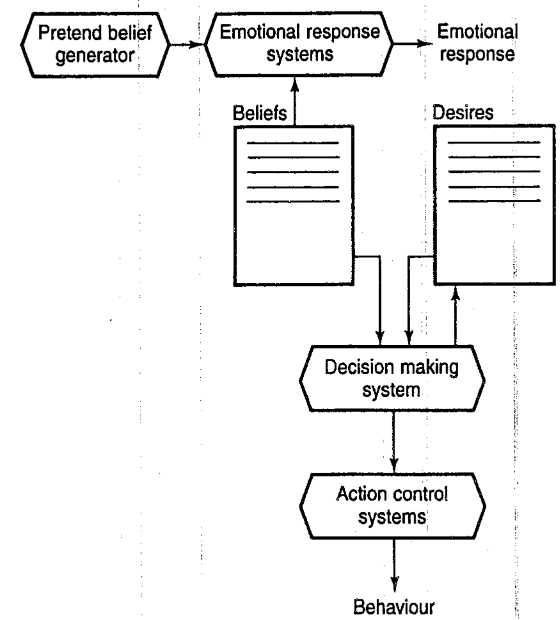

Figure 4.3 Simulation-based account of empathy

Information-based empathy. An alternative view of empathy is that in empathy, as in non-empathic emotion, the emotional response system receives input from the subject's beliefs and memories. Empathic responses might arise when the subject is reminded of events in her past similar to those of the object of empathy. So, for example, if your friend tells you that her dog has died, you might empathise with her via remembering the death of your own dog. Of course, this process of 'remembering' analogous past experiences need not be fully conscious or voluntary. Such 'information-based' accounts of empathy may come in a variety of different forms. According to one plausible version, empathic response arises as follows: a person's beliefs about the situation lead to associations with her own memories of analogous experiences; the emotional response system subsequently takes these memories as inputs and produces the appropriate emotional outputs. This account is sketched in figure 4.4. Unlike off-line simulation, such information-based accounts don't appeal to 'pretend' or deviant inputs. The inputs come from the subject's own knowledge base.

The literature on empathy often appeals to such information-based

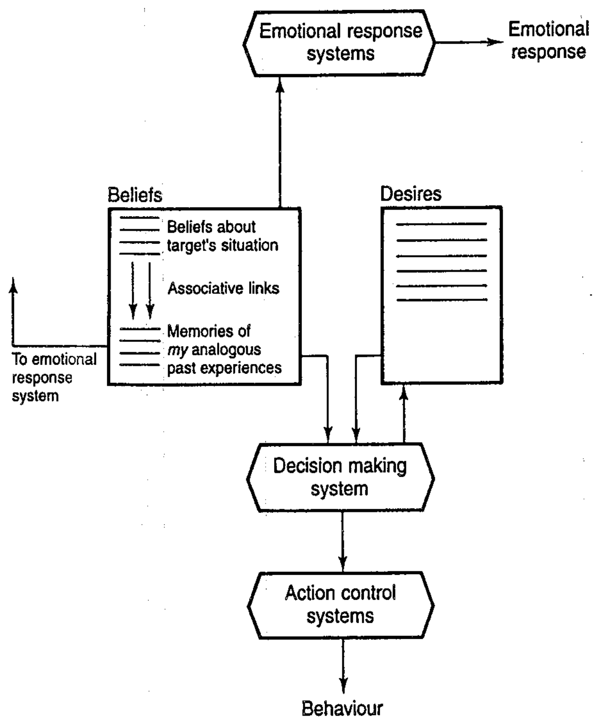

Figure 4.4 Information-based account of empathy

accounts to explain empathic phenomena. Consider, for instance, Hoffman's (1984) 'direct association' mode of empathy:

> When we observe people experiencing an emotion, their facial expression, voice, posture, or any other cue in the situation that reminds us of past situations associated with our experience of that emotion may evoke the emotion in us. The usual example cited is the boy who sees another child cut himself and cry. The sight of the blood, the sound of the cry, or any cue from the victim or the situation that reminds the boy of his own past experience of pain may evoke an empathic distress response (Hoffman, 1984, p. 105)

Eisenberg and Strayer (1987b) also suggest that an information-based account explains a good deal of empathic phenomena. They write, 'It is likely that people often empathise not because they have put themselves cognitively in another's place, but because they have retrieved relevant information from their memories' (Eisenberg and Strayer, 1987b, p. 9). In these passages, neither Hoffman nor Eisenberg and Strayer appeal to pretend input or to mechanisms being taken off line. Rather, they maintain that empathic phenomena can derive from the subject's own information base.

3.2.2 Empathy and off-line simulation

Now that we have sketched both alternatives, we can consider the degree to which off-line simulation captures empathic phenomena. Hoffman (1984) argues that there are at least six different modes of empathy. But rather than try to discuss half a dozen different types of empathy, we'll focus on the three central empathic phenomena discussed by Goldman (1993b): motor mimicry, emotional contagion, and deliberate role taking.

Motor mimicry. Everyday life as well as the laboratory provide numerous instances of people aping the motor behaviour of others. Motor mimicry is familiar from the behaviour of sports spectators. Boxing fans, for example, often bob and weave when watching boxing matches. Infants also display motor mimicry in imitating certain gestures and facial expressions. As noted earlier, a number of researchers claim that motor mimicry is a form of empathy or proto-empathy (Hoffman, 1987, Meltzoff and Gopnik, 1993, p. 358). In keeping with this, Goldman proposes a simulation-based account of mimicry: 'A natural way to explain motor mimicry is in terms of *mental* mimicry: people mentally take the role of another and fail to inhibit (to take 'off line') the behavioural upshot of this mental role taking' (1993b, p. 146).

Apparently, this account would gloss motor mimicry as follows: the subject sees the behaviour, infers the mental states that would produce the behaviour, pretends to have those mental states, and thereby produces homologous behaviour.

Simulation strikes us as quite implausible as a general model of motor mimicry. Performing this kind of simulation would require that a significant portion of folk psychology is in place. In order to perform the relevant simulation, the subject must infer the other's beliefs from her behaviour. The problem with this is that mimicry emerges *very* early. As Goldman himself notes, the capacity for motor mimicry seems to be present at birth (Meltzoff and Moore, 1983). And neither theory-theorists nor simulation theorists think that folk psychology is up and ready at birth.[5] It seems very unlikely that new-borns are capable of inferring beliefs from behaviour. As a result, it's difficult to see how simulation can explain infantile motor mimicry. Andrew Meltzoff, one of the original researchers on infant mimicry, offers a better partial explanation of the phenomenon. He suggests that 'early imitation involves a kind of cross–modal matching. Infants can, at some primitive level, recognise an equivalence between what they see and what they do' (Meltzoff, 1993, p. 222).

Of course, it's possible that the mechanisms for mimicry in infants are different from the mechanisms required for adult motor mimicry. In that case, an advocate of simulation might claim that adult mimicry, unlike mimicry in infants, sometimes does depend on simulation. However, without an argument for why adult motor mimicry is different in kind from infant motor mimicry, any such claim seems ad hoc.

Emotional contagion. Emotional contagion is, as Goldman puts it, 'familiar to all of us through the infectious effects of smiles and laughter' (1993b, p. 142). Emotional contagion is also familiar from the involuntary experience of sorrow on seeing a close friend or relative in grief. Thompson writes, 'Children as well as adults experience the direct, almost involuntary pull of another's emotional expressions in accident settings and other situations eliciting strong affect in others' (1987, p. 124).

Goldman concedes that emotional contagion may not involve role taking.[6] And there is good reason to doubt that emotional contagion derives from simulation. Like motor mimicry, emotional contagion apparently emerges very early. Simner (1971) found that new-borns cry in reaction to neonatal cries significantly more than they cry in reaction to white noise or the cries of an older infant. This has widely been taken to be a primitive form of emotional contagion (Hoffman, 1987; Goldman, 1993b, p. 142). Most infants show further emotional contagion and even more sophisticated empathic behaviour by 18 months (Thompson, 1987; Lamb, 1991).[7] The problem again for the simulation account is that it requires the subject to determine the beliefs of the object of empathy. And it seems unlikely that new-borns are even *capable* of inferring the beliefs of other new-borns. Indeed, we're not sure that *we* can infer the beliefs of crying new-borns.

Again, information-based accounts have no trouble accommodating the data. In reactive new-born cry, as in emotional contagion in general, a subject may be responding on the basis of information retrieved from the subject's own memories. As Hoffman notes, '. . . the sound of the cry, or any cue from the victim or the situation that reminds the boy of his own past experience of pain may evoke an empathic distress response' (Hoffman, 1984, p. 105). On this account, subjects don't have to infer beliefs or feed their emotional response system with pretend inputs. Hence, information-based theories, in contrast to the simulation theory, might explain how new-borns could have the capacity for emotional contagion.

There is another potential problem for the simulation account of emotional contagion. According to simulation theory, the capacity for empathy depends on the capacity for pretence. As a result, since autistic children apparently have a severely impaired capacity for pretence, the simulation theory predicts that they should have a severely impaired capacity for emotional contagion. Autistic children are typically described as being incapable of empathic responses. However, recent experimental evidence suggests that they show far more empathic response than we have been led to expect. Yirmiya *et al.* (1992) showed videotaped stories to both autistic and normal children. They describe the experiment as follows:

> Each one of the [videotaped] segments focuses on the protagonist experiencing one of the following five emotions: happiness, anger, pride, sadness, or fear (e.g., a boy is sad because he lost his dog). After watching each segment, the child is requested to report how he or she feels. . . .
>
> The performance of the autistic children was surprisingly good given the assumptions about the characteristics of autistic individuals. These children were able to give examples of feeling states from their own experience, and many of them showed considerable ability to label the emotions of others, to take the role and perspective of others, and to respond empathetically to the feelings of others (Yirmiya *et al.*, pp. 153, 156–7)

According to the simulation account of emotional contagion, we should expect autistic children to be largely unable to respond empathetically. So, the above results pose a serious challenge to the simulation account. However, autistic children did do less well than the normal children. And this is only one study. So this hardly counts as a decisive argument against the simulation account of emotional contagion. But it does mark out an area of research of considerable importance to the evaluation of simulation theories.

Deliberate role taking. The early onset of both motor mimicry and emotional contagion count as serious prima facie evidence against a simulation account of these phenomena. Hence, there seem to be significant limits to the extent to which off-line, or input-deviant, simulation can explain empathic phenomena. But at the same time, nearly everyone thinks that empathy is sometimes mediated by role taking. For instance, we sometimes *deliberately* try to take another person's perspective. Hoffman claims that the process of putting oneself in another's place, 'being deliberate, may be relatively infrequent – for example, it may be used by parents and therapists who believe they can be more effective if they experience some of their child's or patient's feelings' (Hoffman, 1987, p. 49). Perhaps then, simulation is required for these empathic phenomena. Goldman certainly seems to think that such cases of deliberate role taking establish the simulation account of empathy. He writes:

> the significance of simulation, or role taking, is . . . well established in other areas, e.g., in empathic arousal. In a study by Stotland (1969), subjects were instructed to imagine how they would feel and what sensations they would have in their hands if they were exposed to the same painful heat treatment that was being applied to another person. These subjects gave more evidence of empathic distress than (1) subjects instructed to attend closely to the other person's physical movements and (2) subjects instructed to imagine how the other person felt when he or she was undergoing the treatment (1993b, p. 95)

Goldman apparently maintains that the effects of deliberate role taking are proof that these empathic phenomena derive from simulation. Indeed, Goldman seems to identify role taking with simulation.

It does seem clear that simulation provides a possible model of deliberate role taking. If Bill wanted to empathise with John, he might deliberately try to 'put himself in John's place'. According to the simulation account, to accomplish this Bill would pretend to have some of John's beliefs. These pretend beliefs are fed into Bill's emotional response system which will operate on those feigned inputs as it would on normal inputs. The output, the emotional response, would then be similar to John's emotional response.

Although simulation thus provides a plausible account of deliberate role taking, role taking can also be accommodated by information-based accounts. For instance, an information-based account might maintain that if we want to empathise with John by deliberate role taking, we actively (though perhaps not entirely consciously) try to retrieve memories and relevant information that would lead us to feel what John is feeling. For instance, if Mark wants to empathise with John after the death of his dog, Mark can deliberately try to recall the death of his own dog. That memory might lead Mark to feel the way John does. None of this requires feeding pretend inputs into the emotional response system. Rather, on this information-based account of role taking, if you want to share in someone's grief, you should try to remember similar sad events from your own life.[8]

So, simulation isn't the only account of the processes subserving deliberate role taking. Deliberate role taking can also be explained by information-based accounts. Once we see this, data showing the empathic effects of role taking can no longer be blithely adduced as evidence for simulation. The evidence Goldman cites (Stotland 1969) does show that imagining how we would feel in a situation has a discernible effect on empathic distress. But there are different explanations of how this imaginative process transpires. One possibility is simulation: we pretend to believe that we're being exposed to painful heat. Another possibility is that we try to remember similar experiences from our past.

At present, we know of no compelling evidence to decide between these two accounts of deliberate role taking. As a result, it seems to be an open question whether simulation is *ever* implicated in empathy. One way of trying to answer this question would be by testing whether deliberate role taking is cognitively penetrable. If deliberate role taking is subserved by simulation, subjects should arrive at homologous emotional responses even if they have no memories analogous to the other person's experience. According to information-based accounts of role taking, on the other hand, if there are no analogous memories or relevant information, deliberate role taking will not produce homologous emotional responses. So, for example, on the information-based account, if a certain situation leads people to experience sorrow, but the subject has no analogous memories and no knowledge that the event makes people sad, then deliberate role taking will not lead that subject to feel sorrow. However, on the simulation account, that kind of ignorance should have no effect on empathic response.

3.3 *Mental imagery*

Over the last two decades, a huge literature on mental imagery has developed. This literature has largely focused on the descriptivism/pictorialism debate (Pylyshyn, 1973, 1981; Kosslyn, 1981, 1983, 1994). Descriptivists (e.g., Pylyshyn) maintain that mental images are sentence-like structures; pictorialists (e.g., Kosslyn) maintain that mental images are a kind of non-sentence based representation. In an interesting new paper, Gregory Currie (forthcoming) offers a different perspective on the mental imagery debate.[9] He argues that mental imagery derives from off-line simulation. In this section, we will try to clarify Currie's proposal somewhat. But there remains a fundamental question about Currie's proposal – we aren't sure whether Currie means to offer a novel account of imagery or whether he is arguing that existing accounts of imagery really are simulation theories. We will argue that if Currie intends to present a novel account, then his proposal requires much more elaboration, and on a couple of interpretations the account seems extremely implausible. If, however, Currie's claim is that existing accounts of imagery really are simulation accounts, then, since the notion of a simulation account is at best a vague one, we don't think the claim has any clear truth value, nor do we see why it matters one way or the other.

Currie's account of how imagery derives from simulation is largely contained in the following passage:

> Vision is a process we can think of as having certain standard inputs and outputs. The standard inputs are impulses from the optic nerves which are themselves the result of light impinging on the eyes, and the standard outputs are beliefs – perceptual beliefs – about what is in our immediate environment and where it is in relation

> to our bodies. . . . I suggest that an episode of mental imagery runs the visual system off-line, disconnected from standard inputs and outputs. The central processing part of the visual system – that which is ultimately responsible for visual experience – is seized hold of and operated much as it would if it were taking inputs from the periphery.
>
> Having visual images does not cause us to have beliefs about what is in our immediate environment in the way that vision does. But just as the simulation of belief and desire can mirror the causation of affective states by real belief and desire, so having a mental image of something – a dangerous bear, for example – can produce the affective states that seeing a dangerous bear tends to cause (Currie forthcoming, p. 4)

We find this proposal underdescribed and problematic in a number of ways. One comparatively minor problem is that it seems unlikely that affective states would be the only, or even a primary output of imagery. It's easy, for instance, to imagine a dangerous bear without feeling fear or anxiety. That is to say it's easy to imagine a dangerous bear without having the affective states that would be caused by actually seeing a dangerous bear. Although Currie's paper doesn't make it explicit, presumably the standard outputs of image generation are just mental images. Those images in turn *may* lead to affective states.

We have a more serious complaint about the inputs to the visual processor. Currie tells us neither what the inputs are nor where they come from. Perhaps, in analogy with off-line simulation of behaviour prediction, Currie wants to maintain that there is a pretend input generator that feeds into the central visual processor. The resulting account would claim that in image generation, pretend input is fed into the visual processor which then operates on the pretend input as it would on actual visual input; the output of this is a mental image (which might then lead to an affective state). We've tried to make this functional architecture explicit in figure 4.5.[10]

This functional architecture still leaves many crucial questions unanswered, however. In particular, it's not clear what claim Currie is making about imagery and simulation. We're not sure whether the simulation account of imagery is supposed to be a novel theory of mental imagery or whether the claim is rather that certain established accounts of mental imagery *are* simulation accounts. We'll consider these possibilities in turn.

Currie's theory might be considered novel insofar as it appeals to a pretend input device for imagery. Unfortunately, Currie provides no account of the nature of the pretend input mechanism. Suppose, for instance, that I say to you: form the mental image of a polar bear in the tundra. What is the pretend input mechanism supposed to provide? If the visual processor is to work 'much as it would if it were taking inputs from the periphery', then one might expect that the inputs must be much *like* the

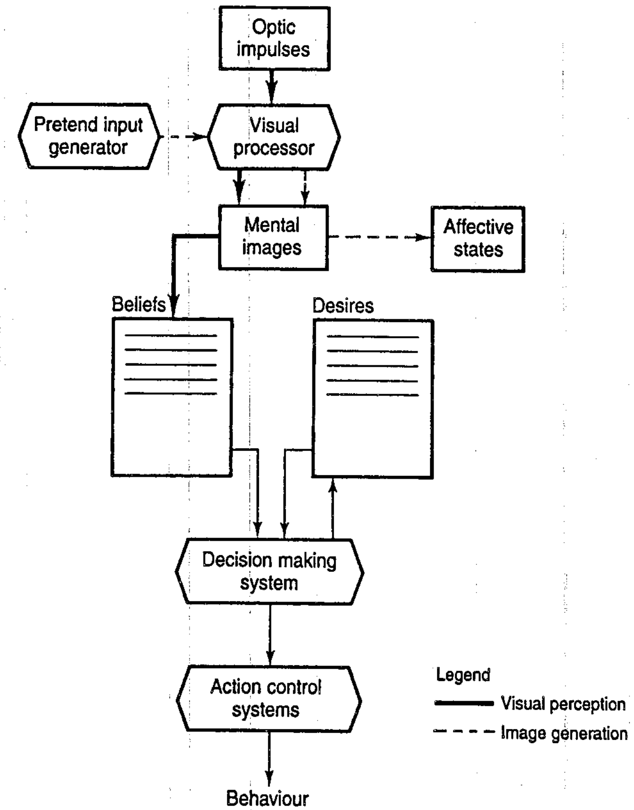

Figure 4.5 Simulation-based account of imagery

inputs from the periphery. And those, as Currie notes, are impulses from the optic nerve. Following this line of thought, the pretend input mechanism would have to take input in the form of linguistic descriptions and produce a reasonable approximation to the signals the optic nerve would produce were we actually to see such a scene. This would provide the simulationist with a novel account of imagery, but at the cost of a mechanism that stretches the bounds of credulity, to put it mildly.

Another possible interpretation of Currie's view is that mental imagery depends on the same general capacity for pretence that simulationists claim is implicated in behaviour prediction. This may be Currie's point in the following passage:

> In my view, the pretend beliefs and desires we have during simulation are no more real beliefs and desires than the simulated fighting on stage is real fighting. We might use the generic term 'imaginings' to cover these pretend states . . . That will help keep track of the commonality between simulating attitudes and simulating vision; having mental images is (a special kind of) imagining seeing (forthcoming, p. 7)

We're not at all sure we understand the supposed commonality between simulating attitudes and simulating vision.[11] In particular, we're not sure whether the pretend inputs for imagery are supposed to be pretend beliefs that might also be fed into the decision maker. It would seem to be somewhat surprising if the visual processor could take the same type of input as the decision maker. In addition, evidence from autism poses a serious *prima facie* problem for the claim that imagery and behaviour prediction depend on the same pretence mechanism. As mentioned earlier, simulationists claim that the reason autistic children fail behaviour prediction tasks is because of a gross deficiency in their ability for pretence (e.g., Currie, 1995; Gordon and Barker 1994). So, if mental imagery depended on this same capacity for pretence, simulationists would predict that autistic children should fail imagery tasks as well (Currie, forthcoming, p. 6). However, there is evidence that autistic children perform well on standard imagery tasks. For instance, they perform at normal levels on Shepard's rotation tasks (Shah, 1988). As a result, simulationists apparently can't appeal to a common pretence mechanism for behaviour prediction and mental imagery.

A final interpretation of Currie's appeal to pretence or imagination is that some 'special kind' of imagining is essential to mental imagery. The nature of this special kind of imagining is, however, left entirely unclear. We still don't know what is involved in 'imagining' to see a polar bear. As a result, it's impossible to say whether appealing to this kind of imagination represents a novel approach to imagery input. For instance, Martha Farah (1984) argues that there must be an image generation mechanism that is independent from the mechanisms of visual perception, and it's not clear how or whether Currie's special kind of imagining differs from Farah's image generation mechanism.

Perhaps, then, Currie doesn't mean to offer a novel theory of imagery. Rather, maybe the point is to argue that prevailing accounts of mental imagery really are simulation accounts. In the recent literature on imagery, there is a virtual consensus that imagery and perception share mechanisms. It's clear that Currie concurs with this. Perhaps Currie's claim is just that insofar as imagery and perception share mechanisms, imagery is a simulation-based capacity. It's difficult to see why that should be the case, though. For consider that it's quite possible that our capacity to predict and explain people's behaviour (the 'folk psychology capacity') and our capacity to predict and explain the behaviour of middle sized objects (the 'folk physics capacity') each depend on information bases (or tacit theories) which share the same inference mechanisms. But surely that's no argument that one of them must be a simulation-based capacity.

There is another established view in the literature that perhaps bears a greater resemblance to simulation theories. Kosslyn and Farah, among others, maintain that while image generation depends on an information base, there is no information base that is *devoted* to image generation. Rather, they maintain that image generation depends on a body of information that is also exploited for visual recognition.

Is Kosslyn's theory plausibly a simulation theory then? Unfortunately, there is no easy way to answer this question since the notion of a simulation theory is far from being well defined. At best it is a family resemblance notion, and one of the anchoring prototypes is the off-line account of behaviour prediction. Kosslyn's view on mental imagery has some important similarities to the off-line account of behaviour prediction. Imagery depends on mechanisms used for vision, and there is no separate information base devoted to imagery. However, there are also striking differences between Kosslyn's account of imagery and the off-line simulation account of behaviour prediction. For instance, according to Gordon and Goldman, behaviour prediction typically doesn't depend on a body of information at all. However, Kosslyn's view, as well as every other established account of imagery, does appeal to a body of information for image generation. Further, on the off-line account of behaviour prediction, the output of the practical reasoner is a proper decision, though it is not fed into the mechanism decisions are usually fed into. But in the case of imagery, the output of the imagery mechanism (whatever that mechanism is like) is considerably different from the outputs of perception. In perception, the output is either visual percepts or beliefs about the environment. But mental images are neither of these. They are not at all as rich as visual percepts. And they typically don't lead to beliefs about the environment. All they seem to do is produce beliefs about the image itself (which may, in turn, produce affective states in some cases). So, unlike the behaviour prediction case, the outputs of imagery aren't normal outputs that are taken off line.

There are, then, significant differences even between Kosslyn's account of imagery and the simulation account of behaviour prediction. Still, we have no in-principle objection to simulationists attaching their label to an established view. However, we are curious about why they should bother. One possible motivation might be to provide support for the simulation account of behaviour prediction. Currie writes, 'If the simulationist can point to other plausible cases of simulation then the claim that our access to other minds is through simulation will sound less like special pleading . . .' (forthcoming, p. 2). We find this sort of inductive argument quite puzzling. For a simulation account of image generation would presumably depend on quite different cognitive mechanisms than the simulation account of behaviour prediction. Even if the visual system is taken off line and fed deviant inputs in the process that leads to imagery, it's unclear why that would give us any reason for thinking that the *decision making system* is taken off line and fed deviant inputs in the process that leads to predictions of other people's behaviour. Further, in the previous paragraph we pointed out additional differences between accounts of mental imagery and the simulation account of behaviour prediction; these differences make it even more implausible that prevailing accounts of mental imagery can provide support for the off-line account of behaviour prediction.

The idea that simulation theory might contribute to our understanding of mental imagery is intriguing. There do seem to be some similarities between accounts of mental imagery and simulation theories. However, as far as we can tell, Currie hasn't offered a plausible simulation theory that provides a genuine alternative to prevailing accounts of imagery. And if that wasn't his intention, then we're not sure how simulation is supposed to illuminate the issue of mental imagery.

Conclusion

Throughout this paper, we've tried to emphasise the extensive explanatory potential of off-line simulation. Off-line simulation promises to offer alternatives to traditional information-based accounts of cognitive capacities. In fact, there already are off-line accounts of a number of radically different capacities, and we've tried to clarify the debate over some of these accounts. In light of the experiments reported here, the off-line simulation account of behaviour prediction still strikes us as quite unpromising. However, this need not diminish the importance of the general strategy of off-line simulation accounts. Indeed, while we are extremely sceptical of the off-line simulation account of behaviour prediction, we think that off-line simulation accounts of counterfactual reasoning, empathy, mental imagery and other phenomena clearly merit further exploration.

NOTES

1 Figure 1 is the same as figure 3 in Stich and Nichols (1992).
2 The metaphor comes from Stephen Schiffer (1981).

3 It's interesting to note that Halle and Stevens offer an argument that closely resembles a familiar argument from Gordon (1986) and Goldman (1989). Gordon and Goldman claim that the theory-theory of behaviour prediction is informationally lavish whereas the simulation alternative is informationally frugal. Similarly, in proposing their 'off-line simulation' account of phoneme recognition, Halle and Stevens argue that the most obvious information-based account of phoneme recognition would be required to posit an utterly enormous information base: 'The size of the dictionary in such an analyser increases very rapidly with the number of admissible outputs, since a given phoneme sequence can give rise to a large number of distinct acoustic outputs. In a device whose capabilities would even remotely approach those of a normal human listener, the size of the dictionary would, therefore, be so large as to rule out this approach' (1962, p. 607).
4 Given the interdisciplinary nature of this volume, perhaps it bears mentioning that counterfactual conditionals are not just conditionals with false antecedents. For material conditionals can have false antecedents; it simply makes the conditional vacuously true, as in, 'If astrology is right, then porcupines make good bed partners.' Counterfactual conditionals, on the other hand, while they typically have false antecedents, make substantive claims which may well be false. For example, the following counterfactual is presumably false: 'If astrology were right, porcupines would make good bed partners.' There is also a grammatical distinction – in English, counterfactual conditionals, unlike material conditionals, are typically expressed in the subjunctive mood.
5 Jerry Fodor (1987, 1992) and Alan Leslie (Leslie and Thais, 1992) have suggested that folk psychology is innate. But even on these views, folk psychology must be triggered. And neither Fodor nor Leslie has suggested that folk psychology is triggered before the child leaves the delivery room.
6 Goldman indicates that insofar as emotional contagion doesn't involve perspective taking, it isn't a clear case of empathy (1993, p. 143). We're not overly concerned with the terminological issues. (We're not concerned at all.) But we do think that emotional contagion is one of the most pervasive and salient cases of 'vicarious sharing of affect'. As such, it's important to determine whether it is better explained by off-line simulation or information-based accounts.
7 Thompson (1987) reports studies in which infants seem to display vicarious responses to the emotional states of others. Thompson writes, 'Among even the youngest children in the sample, the distress of others elicited orienting and, in nearly one third of the 10- to 14-month-olds, distress crying (Zahn-Waxler and Radke-Yarrow, 1982)' (Thompson 1987, p. 131). Thompson takes these and other data to indicate 'a *capacity* for empathy' develops by 18 months (p. 135).
8 Indeed actors claim that when a part calls for them to cry, they sometimes achieve this by focusing on sad memories from their own life.
9 We'd like to think Professor Currie for providing us with his unpublished paper and for allowing us to quote from it.
10 Actually, Currie's account is a bit more complicated than this as he suggests that imagery can also be generated from tacit knowledge (1994b, p. 15). We don't discuss this aspect of Currie's theory since it isn't an off-line simulation account. But we can't resist pointing out how unparsimonious his account is. He posits two radically different mechanisms (simulation and tacit knowledge) each of

which produces mental images. Perhaps imagery really *is* this fractured, but Currie provides no evidence to suggest that it is.

11 One curious feature of Currie's statement is perhaps worth mentioning. Currie writes, 'having mental images is (a special kind of) imagining seeing'. But *having* mental images is, it seems, most naturally read as: experiencing the *output* of the mental image producing process. So here he seems to be saying that the imagining, which is analogous to pretend belief, is the *output*, not the *input*. Perhaps what Currie meant to say was that having mental images *depends on* a special kind of imagining. However, this aspect of the proposal is so vague that we're really not sure what Currie means.

5 Simulation, theory, and content

Jane Heal

1 Introduction

Some, the theory-theorists, say that when we make judgements about the psychological states of others and use such judgements to predict or explain we employ some theory about the psychological. But others, the simulationists, say that we possess no such theory, or at least none complete enough to underpin all our competence with psychological notions; rather, they say, what we do in such situations is simulate others' mental states and processes in ourselves and thus get insight into what others are likely to do.

My aim in this paper is first to offer an argument in favour of simulationism but second to suggest possible limits to the simulationist strategy. I shall suggest that simulation must be central as far as dealing with the contents of others' mental states is concerned but is much less clearly of relevance in dealing with non-content. Thus philosophers and psychologists should not oppose simulation to theory, but should rather ask what is the appropriate realm of each and how they interact.[1]

The topic throughout is the nature of the fully developed adult competence with psychological notions, in the context of predicting others' future psychological states and actions on the basis of knowledge about their current psychological states. I shall not discuss the (it seems to me) importantly different question of how we arrive at judgements about other's thoughts, feelings etc. from knowledge of placement in the environment or bodily behaviour. Also I am not concerned here with the issue of what psychological concepts are and what it is to have possession of them. And finally I shall not touch at all on developmental issues or questions of how children's competence with psychological language grows and changes. I believe that there are implications for all these questions in the considerations which follow, but I shall not pursue them here.[2]

In more detail the structure of what follows is this. Section 2 offers some further clarification of three central notions, simulation, theory, and content, and some remarks on why simulation is at least an option. Section 3 reminds us of some important facts about thought. Section 4 builds on

Theories of theories of mind

edited by

Peter Carruthers
Professor of Philosophy and Director, Hang Seng Centre for Cognitive Studies, University of Sheffield

and

Peter K. Smith
Professor of Psychology, University of Sheffield

Published in association with the Hang Seng Centre for Cognitive Studies, University of Sheffield

CAMBRIDGE
UNIVERSITY PRESS