

- H. K. Wettstein (eds), *Midwest Studies in Philosophy*, 15: *Philosophy of the Human Sciences*, 256–79.
- Morton, A. 1980: *Frames of Mind*. Oxford: Oxford University Press.
- Nisbett, R. E. and Ross, L. 1980: *Human Inference: Strategies and Shortcomings of Social Judgment*. Englewood Cliffs: Prentice-Hall.
- Nozick, R. 1981: *Philosophical Explanations*. Cambridge, Mass.: Harvard University Press.
- Perner, J. 1991: *Understanding the Representational Mind*. Cambridge, Mass.: MIT Press.
- Ripstein, A. 1987: Explanation and empathy. *Review of Metaphysics*, 40, 465–82.
- Stich, S. 1983: *From Folk Psychology to Cognitive Science: The Case Against Belief*. Cambridge, Mass.: MIT Press.
- Strawson, P. F. 1985: Causation and explanation. In B. Vermazen and M. B. Hintikka (eds), *Essays on Davidson: Actions and Events*. Oxford: Oxford University Press, 115–35.
- Vendler, Z. 1984: *The Matter of Minds*. Oxford: Oxford University Press.
- Wilson, E. O. 1975: *Sociobiology*. Cambridge, Mass.: Harvard University Press.

5

Folk Psychology: Simulation or Tacit Theory?

STEPHEN STICH AND SHAUN NICHOLS

1 Introduction

A central goal of contemporary cognitive science is the explanation of cognitive abilities or capacities (Cummins, 1983). During the last three decades a wide range of cognitive capacities have been subjected to careful empirical scrutiny. The adult's ability to produce and comprehend natural language sentences and the child's capacity to acquire a natural language were among the first to be explored (Chomsky, 1965; Fodor, Bever and Garrett, 1974; Pinker, 1989). There is also a rich literature on the ability to solve mathematical problems (Greeno, 1983), the ability to recognize objects visually (Rock, 1983; Gregory, 1970; Marr, 1982), the ability to manipulate and predict the behavior of middle-sized physical objects (McClosky, 1983; Hayes, 1985), and a host of others.

In all of this work, the dominant explanatory strategy proceeds by positing an internally represented 'knowledge structure' – typically a body of rules or principles or propositions – which serves to guide the execution of the capacity to be explained. These rules or principles or propositions are often described as the agent's 'theory' of the domain in question. In some cases, the theory may be partly accessible to consciousness; the agent can tell us some of the rules or principles he is using. More often, however, the agent has no conscious access to the knowledge guiding his behavior. The theory is 'tacit' (Chomsky, 1965) or 'sub-doxastic' (Stich, 1978). Perhaps the earliest philosophical account of this explanatory strategy is set out in Jerry Fodor's paper, 'The Appeal to Tacit Knowledge in Psychological Explanation' (Fodor, 1968). Since then, the idea has been elaborated by Dennett (1978a), Lycan (1981; 1988), and a host of others.

Among the many cognitive capacities that people manifest, there is one cluster that holds a particular fascination for philosophers. Included in this cluster is the ability to *describe* people and their behavior (including their

linguistic behavior) in *intentional terms* – or to ‘interpret’ them, as philosophers sometimes say. We exercise this ability when we describe John as *believing that the mail has come*, or when we say that Anna *wants to go to the library*. By exploiting these intentional descriptions, people are able to offer explanations of each other’s behavior (Susan left the building *because* she believed that it was on fire) and to *predict* each other’s behavior, often with impressive accuracy. Since the dominant strategy for explaining any cognitive capacity is to posit an internally represented theory, it is not surprising that in this area, too, it is generally assumed that a theory is being invoked (Churchland, 1981, 1989; Fodor, 1987; Sellars, 1963; see also Olson et al., 1988). The term ‘folk psychology’ has been widely used as a label for the largely tacit psychological theory that underlies these abilities. During the last decade or so there has been a fair amount of empirical work aimed at describing or modeling folk psychology and tracking its emergence and development in the child (D’Andrade, 1987; Leslie, 1987; Astington et al., 1988).

Recently, however, Robert Gordon, Alvin Goldman and a number of other philosophers have offered a bold challenge to the received view about the cognitive mechanisms underlying our ability to describe, predict and explain people’s behavior (Goldman, 1989; Gordon, 1986; unpublished; Montgomery, 1987; Ripstein, 1987; Heal, 1986).¹ Though they differ on the details, these philosophers agree in denying that an internally represented folk-psychological theory plays a central role in the exercise of these abilities. They also agree that a special sort of mental *simulation* in which we use ourselves as a model for the person we are describing or predicting, will play an important role in the correct account of the mechanisms subserving these abilities. In this paper, although we will occasionally mention the view of other advocates of simulation, our principal focus will be on Gordon and Goldman.

If these philosophers are right, two enormously important consequences will follow. First, of course, the dominant explanatory strategy in cognitive science, the strategy that appeals to internally represented knowledge structures, will be shown to be mistaken in at least one crucial corner of our mental lives. And if it is mistaken there, then perhaps theorists exploring other cognitive capacities can no longer simply take the strategy for granted.

To explain the second consequence we will need a quick review of one of the central debates in recent philosophy of mind. The issue in the debate is the very existence of the intentional mental states that are appealed to in our ordinary explanations of behavior – states like believing, desiring, thinking, hoping, and the rest. *Eliminativists* maintain that there really are no such things. Beliefs and desires are like phlogiston, caloric and witches; they are the mistaken posits of a radically false theory. The theory in question is ‘folk psychology’ – the collection of psychological principles and generalizations which, according to eliminativists (and most of their opponents) underlies our everyday explanations of behavior. The central premise in the eliminativist’s argument is that neuroscience (or connectionism or cognitive science) is on the verge of demonstrating persuasively that folk psychology is false. But if Gordon and Goldman are right, they will have pulled the rug out from under

the eliminativists. For if what underlies our ordinary explanatory practice is not a theory at all, then obviously it cannot be a radically false theory. There is a certain delightful irony in the Gordon/Goldman attack on eliminativism. Indeed, one might almost view it as attempting a bit of philosophical jujitsu. The eliminativists claim that there are no such things as beliefs and desires because the folk psychology that posits them is a radically false theory. Gordon and Goldman claim that the theory which posits a tacitly known folk psychology is *itself* radically false, since there are much better ways of explaining people’s abilities to interpret and predict behavior. Thus, if Gordon and Goldman are right, *there is no such thing as folk psychology!* (Gordon, ch. 2, p. 71; Goldman, ch. 3, p. 93.)

There can be no doubt that if Gordon and Goldman are right, then the impact on both cognitive science and the philosophy of mind will be considerable. But it is a lot easier to doubt that their views about mental simulation are defensible. The remainder of this paper will be devoted to developing these doubts. Here’s the game plan for the pages to follow. In sections 2 and 3, we will try to get as clear as we can on what the simulation theorists claim. We’ll begin, in section 2, with an account of the special sort of simulation that lies at the heart of the Gordon/Goldman proposal. In that section our focus will be on the way that simulation might be used in the *prediction* of behavior. In section 3, we’ll explore the ways in which mental simulation might be used to explain the other two cognitive capacities that have been of special interest to philosophers: *explaining* behavior and producing *intentional descriptions* or *interpretations*. We’ll also consider the possibility that simulation might be used in explaining the *meaning* of intentional terms like ‘believes’ and ‘desires’. Since the accounts of simulation that Gordon and Goldman have offered have been a bit sketchy, there will be a lot of filling in to do in sections 2 and 3. But throughout both sections, our goal will be sympathetic interpretation; we’ve tried hard not to build straw men. In the following two sections, our stance turns critical. In section 4, we will do our best to assemble all the arguments offered by Gordon and Goldman in support of their simulation theory, and to explain why none of them are convincing. In section 5 we will offer two arguments of our own, aimed at showing why, in light of currently available evidence, the simulation theory is very implausible indeed. Section 6 is a brief conclusion.

2 Predicting Behavior: Theory, Simulation and Imagination

Suppose that you are an aeronautical engineer and that you want to predict how a newly built plane will behave at a certain speed. There are two rather different ways in which you might proceed. One way is to sit down with pencil and paper, a detailed set of specifications of the plane, and a state of the art textbook on aerodynamic theory, and try to calculate what the theory entails about the behavior of the plane. Alternatively, you could build a model of the plane, put it in a wind tunnel, and observe how it behaves. You have to

use a bit of theory in this second strategy, of course, since you have to have some idea which properties of the plane you want to duplicate in your model. But there is a clear sense in which a theory is playing the central role in the first prediction and a model or simulation is playing a central role in the second.²

Much the same story could be told if what you want to do is predict the behavior of a person. Suppose, for example, you want to predict what a certain rising young political figure would do if someone in authority tells him to administer painful electric shocks to a person strapped in a chair in the next room. One approach is to gather as much data as you can about the history and personality of the politician and then consult the best theory available on the determinants of behavior under such circumstances. Another approach is to set up a Milgram-style experiment and observe how some other people behave. Naturally, it would be a good idea to find experimental subjects who are psychologically similar to the political figure whose behavior you are trying to predict. Here, as before, theory plays a central role in the first prediction, while a simulation plays a central role in the second.

In both the aeronautical case and the psychological case, we have been supposing that much of the predicting process is carried on outside the predictor. You do your calculations on a piece of paper; your simulations are done in wind tunnels or laboratories. But, of course, it will often be possible to internalize this process. The case is clearest when a theory is being used. Rather than looking in a textbook, you could memorize the theory, and rather than doing the calculations on a piece of paper, you could do them in your head. Moreover, it seems entirely possible that you could learn the theory so well that you are hardly conscious of using it or of doing any explicit calculation or reasoning. Indeed, this, near enough, is the standard story about a wide variety of cognitive capacities.

A parallel story might be told for predictions using simulations. Rather than building a model and putting it in a wind tunnel, you could *imagine* the model in the wind tunnel and see how your imaginary model behaves. Similarly, you could *imagine* putting someone in a Milgram-style laboratory and see how your imaginary subject behaves. But obviously there is a problem lurking here. For while it is certainly possible to imagine a plane in a wind tunnel, it is not at all clear how you could successfully imagine the behavior of the plane unless you had a fair amount of detailed information about the behavior of planes in situations like this one. When the simulation uses a real model of planes, the world tells you how the model will behave. You just have to look and see. But when you are only imagining the simulation, there is no real model for you to look at. So it seems that you must have an internalized knowledge structure to guide your imagination. The theory or knowledge structure that you are exploiting may, of course, be a tacit one, and you may be quite unaware that you are using it. But unless we suppose your imagination is guided by some systematic body of information about the behavior of planes in situations like this one, the success of your prediction would be magic.

When you are imagining the behavior of a person, however, there are various ways in which the underlying system might work. One possibility is that imagining the behavior of a person is entirely parallel to imagining the behavior of a plane. In both cases your imagination is guided by a largely tacit theory or knowledge structure. But there is also a very different mechanism that might be used. In the plane case, you don't have a real plane to observe, so you have to rely on some stored information about planes. You do, however, have a real, human cognitive system to observe – your own. Here's a plausible, though obviously over-simplified, story about how that system normally works:

At any given time you have a large store of beliefs and desires. Some of the beliefs are derived from perception, others from inference. Some of the desires (like the desire to get a drink) arise from systems monitoring bodily states, others (like the desire to go into the kitchen) are 'sub-goals' generated by the decision-making (or 'practical reasoning') system. The decision-making system, which takes your beliefs and desires as input, does more than generate sub-goals, it also somehow or other comes up with a decision about what to do. That decision is then passed on to the 'action controllers' – the mental mechanisms responsible for sequencing and coordinating the behavior necessary to carry out the decision. (Rendered boxologically, the account just sketched appears in figure 5.1.)

Now suppose that it is possible to take the decision making system 'off-line' by disengaging the connection between the system and the action controllers. You might then use it to generate decisions that you are not about to act on. Suppose further that in this off-line mode, you can feed the decision-making system some hypothetical or 'pretend' beliefs and desires – beliefs and desires that you do not actually have, but that the person whose behavior you're trying to predict does. If all this were possible, you could then sit back and let the system generate a decision. Moreover, if your decision-making system is similar to the one in the person whose behavior you're trying to predict, and if the hypothetical beliefs and desires you've fed into your system off-line are close to the ones that he has, then the decision that your system generates will often be similar to the one that his system generates. There is no need for a special internalized knowledge structure here; no tacit folk-psychological theory is being used. Rather, you are using (part of) your own cognitive mechanism as a model for (part of) his. Moreover, just as in the case where the prediction exploits a theory, this whole process may be largely unconscious. It may be that all you are aware of is the prediction itself. Alternatively, if you consciously imagine what the target of your prediction will do, it could well be the case that your imagination is guided by this simulation rather than by some internally represented psychological theory.

We now have at least the outline of an account of how mental simulation might be used in predicting another person's behavior. An entirely parallel

use of simulation in prediction. Here's a passage from his 1986 paper (this volume, p. 70):

[O]ur decision-making or practical reasoning system gets partially disengaged from its 'natural' inputs and fed instead with suppositions and images (or their 'subpersonal' or 'sub-doxastic' counterparts). Given these artificial pretend inputs the system then 'makes up its mind' what to do. Since the system is being run off-line, as it were, disengaged also from its natural output systems, its 'decision' isn't actually executed but rather ends up as an anticipation ... of the other's behavior.

And another, this time from an unpublished manuscript contrasting his view to Fodor's:

The Simulation Theory as I present it holds that we explain and predict behavior not by applying a theory but simply by exercising a skill that has two components: the capacity for practical reasoning – roughly, for making decisions on the basis of facts and values – and the capacity to introduce 'pretend' facts and values into one's decision-making typically to adjust for relevant differences in situation and past behavior. One predicts what the other will decide to do by making a decision oneself – a 'pretend' decision, of course, made only in imagination – after making such adjustments. (Gordon, unpublished, p. 3)

Gordon later suggests that the capacity to simulate in this way may be largely innate:

[Evidence] suggests that the readiness for simulation is a prepackaged 'module' called upon automatically in the perception of other human beings.^[4] It suggests also that supporting and complementing the conscious, reportable procedure we call putting ourselves in the other's place, those neural systems that are responsible for the formation of emotions and intentions are, often without our knowledge, allowed to run off-line: They are partially disengaged from their 'natural' inputs from perception and memory and fed artificial pretend inputs; uncoupled also from their natural output systems, they terminate not as intentions and emotions but as anticipations of, or perhaps just unconscious motor adjustments to, the other's intentions, emotions, behavior. (Gordon, unpublished, p. 5)

3 Other Uses for Simulation: Explanation, Interpretation and the Meaning of Intentional Terms

Let's turn, now, to people's ability to offer *intentional* explanations of other people's actions. How might mental simulation be used to account for that

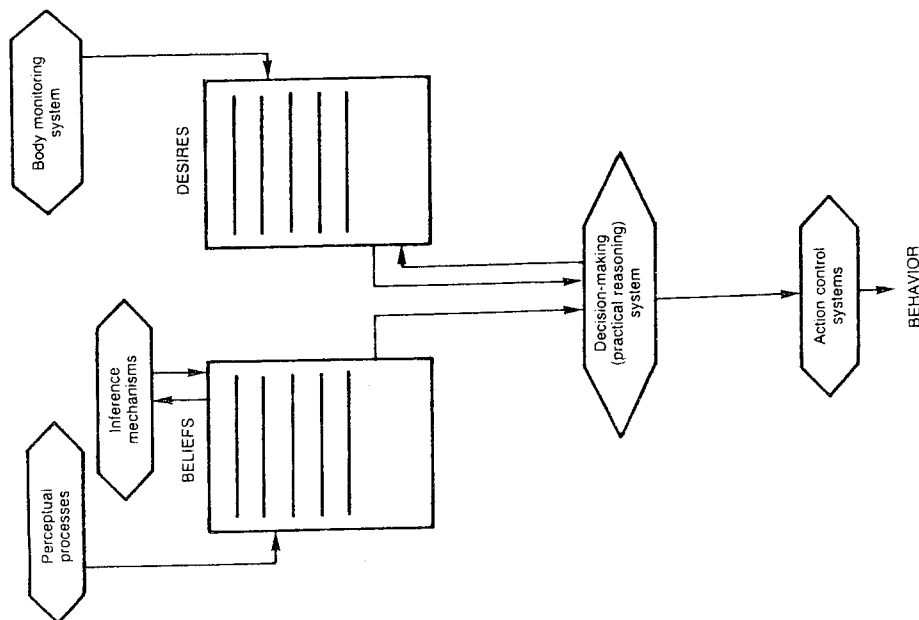


Figure 5.1

story can be told about predicting our own behavior under counterfactual circumstances. If, for example, I want to know what I would do if I believed that there was a burglar in the basement, I can simply take my decision-making system off-line and provide it with the pretend belief that there is a burglar in the basement.³

In the next section we'll try to get clear on how this process of simulation might be used in explaining various other cognitive capacities. But before attending to that task, we would do well to assemble a few quotes to confirm our claim that the story we've told is very close to the one that those we'll be criticizing have in mind. Gordon is much more explicit than Goldman on the

ability? Consider, for example, a case similar to one proposed by Gordon.⁵ We are seated at a restaurant and someone comes up to us and starts speaking to us in a foreign language. How might simulation be exploited in producing an intentional explanation for that behavior?

One proposal, endorsed by both Gordon and Goldman, begins with the fact that simulations can be used in predictions, and goes on to suggest that intentional explanations can be generated by invoking something akin to the strategy of analysis-by-synthesis. In using simulations to predict behavior, hypothetical beliefs and desires are fed into our own decision-making system (being used 'off-line' of course), and we predict that the agent would do what we would decide to do, given those beliefs and desires. A first step in *explaining* a behavioral episode that has already occurred is to see if we can find some hypothetical beliefs and desires which, when fed into our decision mechanism, will produce a decision to perform the behavior we want to explain.

Generally, of course, there will be *lots* of hypothetical beliefs and desires that might lead us to the behavior in question. Here are just a few:

- (a) If we believe someone only speaks a certain foreign language and we want to ask him something, then we would decide to speak to him in that language.
- (b) If we want to impress someone and we believe that speaking in a foreign language will impress him, then we will decide to speak to him in that language.
- (c) If we believe that speaking to someone in a foreign language will make him laugh, and if we want to make him laugh, then we will decide to speak to him in that language.

And so on. Each of these simulation-based predictions provides the kernel for a possible explanation of the behavior we are trying to explain. To decide among these alternative explanations, we must determine which of the input belief/desire pairs is most plausibly attributed to the agent. Some belief/desire pairs will be easy to exclude. Perhaps the agent is a dour fellow; he never wants to make anyone laugh. If we believe this to be the case, then (c) won't be very plausible. In other cases we can use information about the agent's perceptual situation to assess the likelihood of various beliefs. If Mary has just made a rude gesture directly in front of the agent, then it is likely the agent will believe that Mary has insulted him. If the rude gesture was made behind the agent's back, then it is not likely he will believe that she has insulted him. In still other cases, we may have some pre-existing knowledge of the agent's beliefs and desires. But, as both Goldman and Gordon note, it will often be the case that there are lots of alternative explanations that can't be excluded on the basis of evidence about the agent's circumstances or his history. In these cases, Goldman maintains, we simply assume that the agent is psychologically similar to us – we attribute beliefs that are 'natural for us' (Goldman, this volume, p. 90) and reject (or perhaps do not even consider)

hypotheses attributing beliefs that we consider to be less natural (pp. 90–1). Gordon tells much the same story (this volume, p. 65):⁶

No matter how long I go on testing hypotheses, I will not have tried out *all* candidate explanations of the [agent's] behavior. Perhaps some of the unexamined candidates would have done at least as well as the one I settle for, if I settle: perhaps indefinitely many of them would have. But these would be 'far fetched', I say intuitively. Therein I exhibit my inertial bias. The less 'fetching' (or 'stretching', as actors say) I have to do to track the other's behavior, the better. I tend to *feign* only when necessary, only when something in the other's behavior doesn't fit. This inertial bias may be thought of as a 'least effort' principle: the 'principle of least pretending'. It explains why, other things being equal, I will prefer the less radical departure from the 'real' world – i.e. from what I myself take to be the world.⁶

Though the views endorsed by Gordon and Goldman are generally very similar, the two writers do differ in their emphasis. For Gordon, prediction and explanation loom large, while for Goldman, the capacity to *interpret* people, or to describe them in intentional terms, is given pride of place. Part of the story Goldman tells about simulation-based intentional description relies on the account of simulation-based explanation that we have just sketched. One of the ways we determine which beliefs and desires to attribute to people is by observing their behavior and then attributing the intentional states that best explain their behavior. A second simulation-based strategy for determining which beliefs and desires to attribute focuses on the agent's perceptual situation and on his or her 'basic likings or cravings' (Goldman, this volume, p. 82):

From your perceptual situation, I infer that you have certain perceptual experiences or beliefs, the same ones I would have in your situation. I may also assume (pending information to the contrary) that you have the same basic likings that I have: for food, love, warmth, and so on.

As we read them, there is only one important point on which Gordon and Goldman actually *disagree*. The accounts of simulation-based prediction, explanation and interpretation that we have sketched all seem to require that the person doing the simulating must already understand intentional notions like belief and desire. A person can't pretend he believes that the cookies are in the cookie jar unless he understands what it is to believe that the cookies are in the cookie jar; nor can a person imagine that she wants to make her friend laugh unless she understands what it is to want to make someone laugh. Moreover, as Goldman notes, when simulation is used to attribute intentional states to agents, it 'assumes a prior understanding of what state it is that the interpreter attributes to [the agent]' (Goldman, this volume, p. 94). Can the process of

simulation somehow be used to explain the meaning or truth conditions of locations like 'S believes that *p*' and 'S desires that *q*'? Goldman is skeptical, and tells us that 'the simulation theory looks distinctly unpromising on this score' (this volume, p. 93). But Gordon is much more sanguine. Building on earlier suggestions by Quine, Davidson and Stich, he proposes the following account (Gordon, this volume, p. 68):

My suggestion is that

(2) [Smith believes that Dewey won the election]

to be read as saying the same thing as

(1) [Let's do a Smith simulation. Ready? Dewey won the election] though less explicitly.

We are not at all sure we understand this proposal, and Gordon himself concedes that 'the exposition and defense of this account of belief are much in need of further development' (this volume, p. 68). But no matter. We think we do understand the simulation-based accounts of prediction, explanation and interpretation that Gordon and Goldman both endorse. We're also pretty certain that none of these accounts is correct. In the sections that follow, we will try to say why.

4 Arguments in Support of Simulation-based Accounts

In this section we propose to assemble all the arguments we've been able to find in favor of simulation-based accounts and say why we don't think any of them is persuasive. Then, in the following section, we will go on to offer some arguments of our own aimed at showing that there is lots of evidence that simulation-based accounts cannot easily accommodate, though more traditional theory-based accounts can. Before turning to the arguments, however, we would do well to get a bit clearer about the questions that the arguments are (and are not) intended to answer.

The central idea in the accounts offered by Gordon and Goldman is that in predicting, explaining or interpreting other people we simulate them by using part of *our own* cognitive systems 'off-line'. There might, of course, be other kinds of simulation in which we do not exploit our own decision-making system in order to model the person we are simulating. But these other sorts of simulation are not our current concern. To avoid confusion, we will henceforth use the term *off-line simulation* for the sort of simulation that Gordon and Goldman propose. The question in dispute, then, is whether off-line simulation plays a central role in predicting, explaining or interpreting other people. Gordon and Goldman say yes; we say no.

It would appear that the only serious alternatives to the off-line simulation story are various versions of the 'theory-theory' which maintain that prediction, explanation and interpretation exploit an internally represented theory or knowledge structure – a tacitly known 'folk psychology'. So if an

advocate of off-line simulation can mount convincing arguments against the theory-theory, then he can reasonably claim to have made his case. The theory-theory is not the only game in town, but it is the only *other* game in town. It is not surprising, then, that in defending off-line simulation Gordon and Goldman spend a fair amount of time raising objections to the theory-theory.

There are, however, some important distinctions to be drawn among different types of theory-theories. Until fairly recently, most models that aimed at explaining cognitive capacities posited internally represented knowledge structures that invoked explicit rules or explicit sentence-like principles. But during the last decade there has been a growing dissatisfaction with sentence-based and rule-based knowledge structures, and a variety of alternatives have been explored. Perhaps the most widely discussed alternatives are connectionist models in which the knowledge used in making predictions is stored in the connection strengths between the nodes of a network. In many of these systems it is difficult or impossible to view the network as encoding a set of sentences or rules (Ramsey, Stich and Garon, 1990). Other theorists have proposed quite different ways in which non-sentential and non-rule-like strategies could be used to encode information. (See, for example, Johnson-Laird, 1983.)

Unfortunately, there is no terminological consensus in this domain. Some writers prefer to reserve the term 'theory' for sentence-like or rule-based systems. For these writers, most connectionist models do not invoke what they would call an internally represented theory. Other writers are more liberal in their use of 'theory', and are prepared to count just about any internally stored body of information about a given domain as an internally represented theory of that domain. For these writers, connectionist models and other non-sentential models do encode a tacit theory. We don't think there is any substantive issue at stake here. But the terminological disagreements can generate a certain amount of confusion. Thus, for example, someone who used 'theory' in the more restrictive way might well conclude that if a connectionist (or some other non-sentence-based) account of our ability to predict other people's behavior turns out to be the right one, then the theory-theory is mistaken. So far, so good. But it is important to see that the falsity of the theory-theory (narrowly construed) is no comfort at all to the off-line simulation theorist. The choice between off-line simulation theories and theory-theories is plausibly viewed as exhaustive only when 'theory' is used in the *wide* rather than the restrictive way. For the remainder of this paper, we propose to adopt the wide interpretation of 'theory'. Using this terminology, the geography of the options confronting us is represented in figure 5.2.⁷ In the pages that follow, we will be defending option (A) in answer to Question (I). We take no stand at all on Question (II). So much for getting clear on the questions. Now let's turn to the arguments.

Argument 1: No one has been able to state the principles of the internally represented folk-psychological theory posited by the theory-theory.

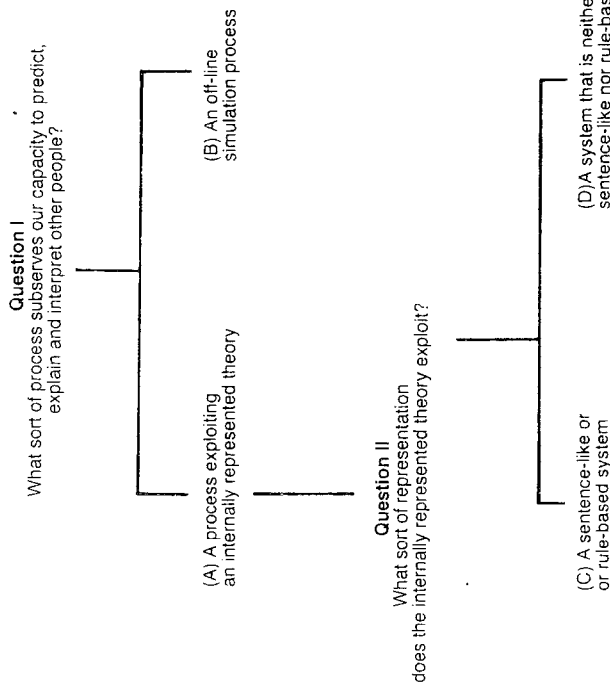


Figure 5.2

Both Goldman and Gordon go on at some length about the fact that it has proven very difficult to state the principles or laws of the folk psychological theory that, according to the theory-theorist, guide our interpretations and predictions (Goldman, 1989, p. 167 of original version):

[A] Attempts by philosophers to articulate the putative laws or 'platitudes' that comprise our folk theory have been notably weak. Actual illustrations of such laws are sparse in number; and when examples are adduced, they commonly suffer from one of two defects: vagueness and inaccuracy . . . But why, one wonders, should it be so difficult to articulate laws if we appeal to them all the time in our interpretative practice? (See also Gordon, this volume, p. 67; and unpublished, sec. 3.7.)

Reply: Goldman is certainly right about one thing. It is indeed very difficult to articulate the principles of folk psychology precisely and accurately. But it is hard to see why this fact should be of any comfort to advocates of the off-line simulation theory. For much the same could be said about the knowledge structures underlying all sorts of cognitive capacities. It has proven enormously difficult to state the principles underlying a speaker's capacity to judge the grammaticality of sentences in his language. Indeed, after three

decades of sustained effort, we don't have a good grammar for even a single natural language. Nor do we have a good account of the principles underlying people's everyday judgements about the behavior of middle-sized physical objects, or about their ability to solve mathematical problems, or about their ability to play chess, etc. But, of course, in all of these domains, the theory-theory really is the only game in town. The off-line simulation story makes no sense as an account of our ability to judge grammaticality, or of our ability to predict the behavior of projectiles.

The difficulties encountered by those who have sought to describe the rules or principles underlying our grammatical (or mathematical or physical) abilities have convinced a growing number of theorists that our knowledge in these domains is not stored in the form of rules or principles. That conviction has been an important motive for the development of connectionist and other sorts of non-sentential and non-rule-based models. But none of this should encourage an advocate of the off-line simulation theory. The dispute between connectionist models and rule-based models is the dispute between (C) and (D) in figure 5.2. And that is a dispute *among theory-theorists*. Of course on a narrow interpretation of 'theory', on which only rule-based and sentence-based models count as theories, the success of connectionism would indeed show that the 'theory-theory' is mistaken. But, as we have taken pains to note, a refutation of the theory-theory will support the off-line simulation account only when 'theory' is interpreted broadly.

Argument 2: Mental simulation models have been used with some success by a number of cognitive scientists.

Here's how Goldman makes the point (this volume, p. 86):

[S]everal 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 condition for a 'run', they say, can either be left at realistic default values or modified to assume some special contingency. Similarly, Rumelhart [et al.] describe the importance of 'mental models' of the world, in particular, models that simulate how the world would respond to one's hypothetical actions.

Reply: Here, again, it is our suspicion that ambiguity between the two interpretations of 'theory' is lurking in the background and leading to mischief. The 'simulation' models that Goldman cites are the sort that would be classified under (D) in figure 5.2. If they are used in the best explanation of a given cognitive capacity, then that capacity is subserved by a tacit theory, and *not* by an off-line simulation. Of course when 'theory' is read narrowly, this sort of

simulation will not count as a tacit theory. But, as already noted, on the narrow reading of 'theory' the falsity of internalized theory accounts lends no support at all to the off-line simulation theory.

Argument 3: 'To apply the alleged common-sense theory would demand anomalous precocity.'

What we've just quoted is a section heading in one of Gordon's unpublished papers.⁸ He goes on to note that recent studies have shown children as young as two and a half 'already see behavior as dependent on belief and desire'. It is, he suggests, more than a bit implausible that children this young could acquire and use 'a theory as complex and sophisticated' as the one that the theory-theory attributes to them. Goldman elaborates the argument as follows (this volume, p. 80):

[C]hildren seem to display interpretive skills by the age of four, five or six. If interpretation is indeed guided by laws of folk psychology, the latter must be known (or believed) by this age. Are such children sophisticated enough to represent such principles? And how, exactly, would they acquire them? One possible mode of acquisition is cultural transmission (e.g. being taught them explicitly by their elders). This is clearly out of the question, though, since only philosophers have even tried to articulate the laws, and most children have no exposure to philosophers. Another possible mode of acquisition is private construction. Each child constructs the generalizations for herself, perhaps taking clues from verbal explanations of behavior which she hears. But if this construction is supposed to occur along the lines of familiar modes of scientific theory construction, some anomalous things must take place. For one thing, all children miraculously construct the same nomological principles. This is what the (folk-) theory theory ostensibly implies, since it imputes a single folk psychology to everyone. In normal cases of hypothesis construction, however, different scientists come up with different theories.

Reply: There is no doubt that if the theory-theory is right, then the child's feat is indeed an impressive one. Moreover, it is implausible to suppose that the swift acquisition of folk psychology is subserved by the same learning mechanism that the child uses to learn history or chemistry or astronomy. But, once again, we find it hard to see how this can be taken as an argument against the theory-theory and in favor of the off-line simulation theory. For there are other cases in which the child's accomplishment is comparably impressive and comparably swift. If contemporary generative grammar is even *close* to being right, the knowledge structures that underlie a child's linguistic ability are enormously complex. Yet children seem to acquire the relevant knowledge structures even more quickly than they acquire their knowledge of folk psychology. Moreover, children in the same linguistic community all acquire much the same grammar, despite being exposed to significantly different

samples of what will become their native language. Less is known about the knowledge structures underlying children's abilities to anticipate the behavior of middle-sized physical objects. But there is every reason to suppose that this 'folk physics' is at least as complex as folk psychology, and that it is acquired with comparable speed. Given the importance of all three knowledge domains, it is plausible to suppose that natural selection has provided the child with lots of help – either in the form of innate knowledge structures or in the form of special-purpose learning mechanisms. But whatever the right story about acquisition turns out to be, it is perfectly clear that in the case of grammar, and in the case of folk physics, what is acquired must be some sort of internally represented theory. Off-line simulation could not possibly account for our skills in those domains. Since the speed of language acquisition and the complexity of the knowledge acquired do not (indeed, could not) support an off-line simulation account of linguistic ability, we fail to see why Gordon and Goldman think that considerations of speed and complexity lend any support at all to the off-line simulation account for our skills in predicting, explaining and interpreting behavior.

Argument 4: The off-line simulation theory is much simpler than the theory-theory.

Other things being equal, we should surely prefer a simple theory to a more complex one. And on Gordon's view (unpublished, sec. 3, p. 7):

the simulation alternative makes [the theory-theory] strikingly unparsimonious. Insofar as the store of causal generalizations posited by [the theory-theory] mirrors the set of rules *our own* thinking typically conforms to, the Simulation Theory renders it altogether otiose. For whatever rules our own thinking typically conforms to, our thinking continues to conform to them within the context of simulation . . . In the light of this far simpler alternative, the hypothesis that people must be endowed with a special stock of laws corresponding to rules of logic and reasoning is unmotivated and unparsimonious.

Reply: When comparing the simplicity of a pair of theories, it is important to look at the whole theory in both cases, not just at isolated parts. It is our contention that if one takes this broader perspective, the greater parsimony of the simulation theory simply disappears. To see the point, note that for both the theory-theory and the simulation theory the mechanism subserving our predictions of other people's behavior must have two components. One of these may be thought of as a data base that somehow stores or embodies information about how people behave. The other component is a mechanism which applies that information to the case at hand – it extracts the relevant facts from the data base. Now if we look only at the data base, it does indeed seem that the theory-theory is 'strikingly unparsimonious' since it must posit an elaborate system of internally represented generalizations or rules – or

perhaps some other format for encoding the regularities of folk psychology. The simulation theory, by contrast, uses the mind's decision-making system as its 'data base', and that decision-making system would have to be there on any theory, because it explains how we make real, 'on-line' decisions. So the off-line simulation theory gets its data base for free.

But now let's consider the other component of the competing theories. Merely *having* a decision-making system will not enable us to make predictions about other people's behavior. We also need the capacity to take that system 'off-line', feed it 'pretend' inputs and interpret its outputs as predictions about how someone else would behave. When we add the required cognitive apparatus, the picture of the mind that emerges is sketched in figure 5.3. Getting this 'control mechanism' to work smoothly is sure to be a *very* non-trivial task. How do things look in the case of the theory-theory? Well, no matter how we go about making predictions about other people, it is clear that in making predictions about physical systems we can't use the off-line simulation strategy: we have to use some sort of internalized theory (though, of course, it need not be a sentence-like or rule-based theory). Thus we know that the mind is going to have to have some mechanism for extracting information from internalized theories and applying it to particular cases. (In figure 5.1 we have assumed that this mechanism is housed along with the other 'inference mechanisms' that are used to extract information from pre-existing beliefs.) If such a mechanism will work for an internally represented folk physics, it is plausible to suppose that, with minor modifications, it will also work for an internally represented folk psychology. So while the simulation theorist gets the data base for free, it looks like the theory-theorist gets the 'control mechanism' for free. All of this is a bit fast and loose, of course. But we don't think either side of this argument can get much more precise until we are presented with up-and-running models to compare. Until then, neither side can gain much advantage by appealing to simplicity.

Argument 5: When we introspect about our predictions of other people's behavior, it sometimes seems that we proceed by imagining how we would behave in their situation.

Here is how Goldman makes the point (this volume, p. 82):

The simulation idea has obvious initial attractions. Introspectively, it seems as if we often try to predict others' behavior – or predict their (mental) choices – by imagining ourselves in their shoes and determining what we would choose to do.

And here is Gordon (this volume, p. 63):

[C]hess players report that, playing against a human opponent or even against a computer, they visualize the board from the other side, taking the opposing pieces for their own and vice versa. Further, they pretend

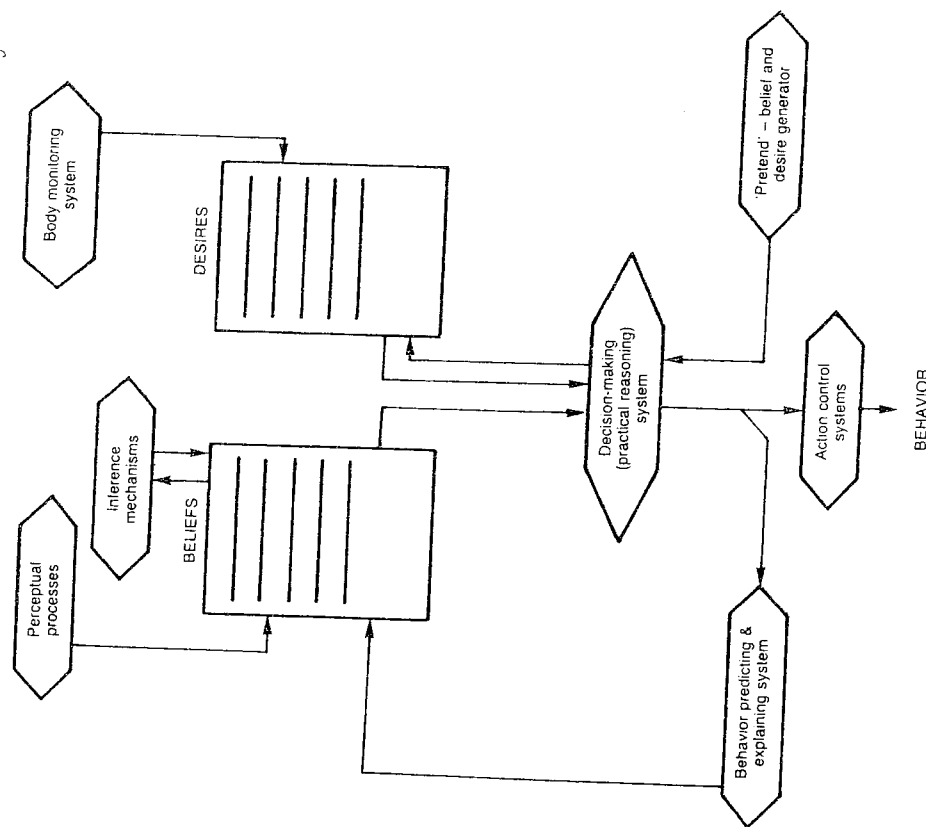


Figure 5.3

that their reasons for action have shifted accordingly ... Thus transported in imagination, they 'make up their mind what to do'.

Both authors are aware that appeal to introspection can be a two-edged sword, since it also often happens that we predict other people's behavior *without* introspecting any imaginary behavior (Goldman, this volume, p. 87):

[T]here is a straightforward challenge to the psychological plausibility of the simulation approach. It is far from obvious, introspectively, that we regularly place ourselves in another person's shoes, and vividly envision what we would do in his circumstances.