In P. Carruthers and P.K. Smith (eds), *Theories of Theories of Mind*. Cambridge: Cambridge University Press, 119–37.

The Mental Simulation Debate: A Progress Report

TONY STONE AND MARTIN DAVIES

1. Introduction

For philosophers, the current phase of the debate with which this volume is concerned can be taken to have begun in 1986, when Jane Heal and Robert Gordon published their seminal papers (Heal, 1986; Gordon, 1986; though see also, for example, Stich, 1981; Dennett, 1981). They raised a dissenting voice against what was becoming a philosophical orthodoxy: that our everyday, or folk, understanding of the mind should be thought of as theoretical. In opposition to this picture, Gordon and Heal argued that we are not theorists but simulators. For psychologists, the debate had begun somewhat earlier when Heider (1958) produced his work on lay psychology; and in more recent times the psychological debate had continued in developmental psychology and in work on animal cognition.

But the debate has a much longer provenance than those datings suggest; for it goes back, at least, to disputes in the eighteenth century about whether the methods that had been so successful in the natural sciences were also appropriate for the human or moral sciences. Today's friends of mental simulation stand in a tradition that includes Vico, Herder, Croce, and particularly Collingwood (1946).

1.1 Nine questions

Given the inter-disciplinary ancestry of the debate, it is no surprise that current discussion ranges over a number of distinguishable — indeed, often fairly independent — questions. We suggest that it is useful to separate out some of these questions, and here we identify nine.

- (a) What is it to have mastery of the mental concepts that are deployed in our folk psychological practice? (The concept mastery question)

- (b) What is the best philosophical account of the kinds of states postulated by the folk when they engage in folk psychological practice? (The metaphysical question)

- (c) What are the key characteristics of our folk psychological practice — particularly, our practices of attribution, explanation and prediction? (The descriptive question about normal adult folk psychological practice)

- (d) What resources do mature adult humans draw upon as they go about the business of attributing mental states, and predicting and explaining one another's mental states and actions? (The explanatory question about normal adult folk psychological practice)

- (e) What information processing mechanisms need to be postulated in order to provide a psychological explanation of the way in which humans actually attribute mental states, and predict and explain one another's actions? (The question of information processing underpinnings of normal adult practice)

- (f) What course of development do human beings follow as they develop the ability to engage in folk psychological practice? (The descriptive question about development)

- (g) What explanatory account is to be given of this course of development? (The explanatory question about development)

- (h) What mechanisms need to be postulated in order to explain the changes that are seen in the child's folk psychological abilities? (The question of mechanisms responsible for change)

- (i) What explanatory theory can we give that explains the deficits and disorders to which the development of folk psychological practice is subject? (The question about developmental disorders)

Given the differences amongst these questions, we should not assume that a position that starts life as an answer to a philosophical question (like (a) or (b)) can be simply transposed into an answer to an empirical question (like (c) – (i)). So, there is no way to be brief over these nine questions. What we shall do, then, is to focus our comments on the explanatory question about normal adult folk psychological practice. Along the way, there will be some mention of the concept mastery question, and of the descriptive question about normal adult practice. The metaphysical question, and the questions about information processing underpinnings and about development, will be left to one side.

2. The theory theory

Here is one way in which a version of the theory theory might be developed. It might be said that the mental concepts that comprise our everyday or folk psychology — such as *belief*, *desire*, *hope*, *being in pain*, and so on — are part of a linked network of mental concepts, so that understanding any one of these concepts requires understanding some or all of the others. Thus, grasping the concept of belief, for example, requires that one has mastered connections between belief and desire of the form (Churchland, 1988, p. 58/9):

> Persons who want that *P*, and believe that *Q* would be sufficient to bring about *P*, and have no conflicting wants or preferred strategies, will try to bring it about that *Q*.

or (Botterill, this volume, p. xxx):

> [Action Principle] An agent will act in such a way as to satisfy, or at least to increase the likelihood of satisfaction of, his/her current strongest desire in the light of his/her beliefs.

Similarly, of course, grasping the concept of desire requires mastering the same kinds of connections.

The argument in favour of this kind of view is well known. Imagine, it is said, someone who claims to have the concept of belief — indeed someone who claims to understand what it is to believe that its being the case that *Q* would be sufficient to bring it about that *P* — but who denies that someone with that belief would try, *ceteris paribus,* to bring it about that *Q* were she to desire that *P*. Then — say those who support this approach to questions about grasp of mental concepts — it is clear that, whatever concept it is that such a person is using, it is not the concept of *belief*. The same point can be made *mutatis mutandis* about grasping the concept of desire, and particularly about understanding the idea of desiring that *P*.

Such arguments for the view that grasp of certain concepts requires commitment to a family of inferential principles have considerable plausibility. This is especially so when grasp of a concept cannot be directly manifested by way of demonstrative identification of objects or happenings as falling under it. And indeed, many advocates of the theory theory would stress that beliefs are unobservable, and so are not clear candidates for demonstrative identification (Fodor, 1987, p. 7).

2.1 The analogy with science

A theory theorist may now take the following step. If our grasp of mental concepts depends upon our mastery of the inferential connections in a network of concepts, then there seems to be an analogy between mental concepts and those scientific concepts that get their sense from the scientific theory in which they are embedded. A concept such as *quark* can only be grasped via knowledge of a theory of sub-atomic physics. As in the case of *belief*, there is no prospect presently of being able to identify quarks demonstratively. Grasp of the concept *quark* depends upon mastery of the theory in which that concept is embedded.

In this spirit, David Lewis (1972) argues that we should see folk psychology 'as a term-introducing scientific theory, though one invented long before there was any such institution as professional science' (p. 256). The theory is formulated in the following way (ibid.):

> Collect all the platitudes you can think of regarding the causal relations of mental states, sensory stimuli and motor responses. . . . Include only platitudes which are common knowledge among us — everyone knows them, everyone knows that everyone else knows them, and so on.

This analogy between mental concepts and scientific concepts is the source for one version of the theory theory of folk psychology — a version that begins as an answer to the concept mastery question (a).

The rather limited analogy with professional science is taken a step further by those who compare the forms of explanation deployed in folk psychological practice with the deductive-nomological explanations that are usually regarded as characteristic of scientific practice. That step positions the theory theorist to return an answer to the explanatory question about normal adult folk psychological practice (d); the answer, namely, that the resources drawn upon in folk psychological practice include knowledge of a psychological theory. But we should note that the theory theorist's answer to the explanatory question rests upon an answer to the descriptive question (c), especially as that question relates to the key characteristics of folk psychological explanation.

Explanations in folk psychology are sometimes said to be 'rationalising' explanations. Philip Pettit (1986, p. 45), for example, describes them as 'normalising' rather than 'regularising' explanations, and John McDowell (1985, p. 389) says that 'the concepts of the propositional attitudes have their proper home in explanations of a special sort' — a sort contrasted with subsumptive or regularising explanation. But the theory theorist's answer to question (d) about the explanation of our folk psychological practice — including our explanatory practice — begins from an answer to question (c) that says that the explanatory practice itself is not radically different in kind from giving explanations in terms of 'how things generally tend to happen' (McDowell, ibid.). Thus, Fodor states that commonsense psychological explanations (1987, p. 7):

> exhibit the 'deductive structure' that is so characteristic of explanation in real science. There are two parts to this: the theory's underlying generalizations are defined over unobservables, and they lead to its predictions by iterating and interacting rather than being directly instantiated.

According to this kind of position, then, folk psychological explanation and scientific explanation are of essentially the same kind — and draw upon the same kinds of resources. As Heal (1986, p. 135) describes the position:

> We are said to view other people as we view stars, clouds or geological formations. People are just complex objects in our environment whose behaviour we wish to anticipate but whose causal innards we cannot perceive. We therefore proceed by observing the intricacies of their external behaviour and formulating some hypotheses about how the insides are structured.

A fortiori, on this view, there is continuity between folk psychology and scientific psychology. Scientific psychological practice is an extension of the way that people actually do go about their everyday explanatory activities — 'but in more detail and with more statistical accuracy' (Heal, ibid.).

This is, of course, a substantive — and even controversial — claim. There is no evident incoherence in combining the idea that, when we are doing science, we can treat human beings as 'complex objects . . . whose behaviour we wish to anticipate' with the thought that ordinary folk psychological activity goes on in a fundamentally different way. Indeed, there would appear to be room for a range of views of this general type, differing over what is distinctive about folk psychological explanation. One variation on the theme would be that suggested by Pettit and McDowell. Another would stress the role of lore — rather than law — in everyday understanding and explanation (see Botterill, this volume; Heal, this volume, for examples). (See also Wolpert, 1992, for a view of science as fundamentally different in kind from commonsense thinking.) We shall return, in the concluding section of this chapter, to the idea that folk psychological explanation is fundamentally different from explanation by subsumption.

Up to this point, the weight carried by the notion of a psychological theory has depended upon an analogy with professional science. But this analogy may be somewhat problematic for the version of the theory theory that we have been considering. After all, in many respects the practice of folk psychology looks utterly unlike that of a professional science — quantum physics, or inorganic chemistry, or molecular biology, or neurophysiology. Folk psychological practice does not bear the marks of a scientific research programme. It is not, for example, written up in learned journals or in text books; it is not subject to rigorous empirical investigation; nor does it have to be actively taught.

There are a number of strategies that could be taken by someone who seeks to articulate a version of the theory theory, but who does not want to become entangled in these disanalogies between folk psychology and professional science. One strategy is simply to focus on the deductive-nomological form of explanation. Another strategy — not unconnected with what Fodor (1987, p. 7) says about 'generalizations [that] lead to . . . predictions by iterating and interacting rather than being directly instantiated' — is to stress the idea of a theory as providing a framework within which particular cases can be systematically related to each other (Heal, this volume). '[T]heories must contain principles that provide a systematic integration of knowledge' (Botterill, this volume, p. xxx).

2.2 The analogy with theoretical linguistics

Some friends of the theory theory may abandon the analogy with professional science altogether, preferring a different analogy with theoretical linguistics. This analogy sees the theory theory as adopting what Stich and Nichols (1992) call 'the dominant explanatory strategy' in cognitive science (1992, pp. 35–6):

> 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 behaviour.

But, just as the analogy with professional science can be problematic, so also this analogy with linguistics raises some delicate questions.

The analogy prompts a host of questions about the notion of tacit knowledge, for example. The theory theorist adopting this strategy needs to give some account of tacit knowledge that makes it clear just what attributions of tacit knowledge amount to — particularly, it must be clear that an attribution of tacit knowledge amounts to more than just a summary description of the practice or ability that is to be explained.

Furthermore, there may be an important disanalogy between tacit knowledge in the case of linguistics and tacit knowledge in the case of folk psychology. In the case of linguistics there is something to be said for the idea that the content of the tacit knowledge does not have to be conceptualised by the subject whose tacit knowledge it is. Ordinary language users do not grasp the concepts of linguistic theory, and so their tacit knowledge of linguistic principles does not constitute their mastery of those concepts. But in the case of folk psychology, ordinary practitioners do possess the concepts that will, presumably, be central in the tacitly known theory. Botterill (this volume) makes the related point that it would be natural for a theory theorist to deny that the principles of the known folk psychological theory are completely inaccessible to consciousness, given that folk psychological practice includes providing explanations: '[W]hy should citing the psychological antecedents of action . . . appear to us to be explanatory, unless we have some sort of awareness of the principles involved?' (p. xxx).

2.3 The 'body of knowledge' strategy

We turn now to a strategy that takes a relaxed view of the disanalogies between folk psychology and the theories that are characteristic of professional science. According to this version of the theory theory, the key idea is that we can account for the explanatory abilities of human beings in the psychological domain by attributing to them possession of a body of knowledge about that domain. The body of knowledge might or might not be best thought of as structured in an especially theoretical way — as formulated in axioms and theorems, for example. But what is crucial is that it is an articulated body of knowledge that is specifically about the domain in question — here, the psychological domain. Stich and Nichols (1992) are the clearest advocates of this more relaxed (they say 'less restrictive') strategy.

The motivation for this strategy is clear. Stich and Nichols are attempting to provide a version of the theory theory that is broad enough in its conception to withstand criticisms launched against specific forms that it might take. They are trying to provide a generic formulation of the theory theory position (1995, p. xxx):

> If this [sc. the analogy with professional science] is correct, then folk psychology will bear a strong resemblance to the standard philosophical portrait of scientific theories in domains like physics and chemistry. But, of course, there are lots of domains of commonsense knowledge in which it is rather implausible to suppose that the mentally represented 'knowledge structure' includes theoretical constructs linked together in lawlike ways. Knowledge of cooking, or of current affairs are likely candidates here, as is the knowledge that underlies our judgements about what is polite and impolite in our culture. And it is entirely possible that folk psychological knowledge will turn out to resemble the knowledge structures underlying cooking or politeness judgements rather than the knowledge structures that underlie the scientific predictions produced by a competent physicist or chemist.

Formulating the theory theory in this more relaxed way has an important tactical role to play in the debate between the theory theory and the simulation theory. As Stich and Nichols point out, taking the theory theory in this way means that (1995, p. xxx):

> . . . even if it could be shown that people do not exploit lawlike

> generalizations in predicting and explaining other people's behavior, this would not show that the theory theory is wrong, and it would not provide any significant degree of support for the simulation theory.

Apart from this tactical advantage, adopting Stich and Nichols's strategy also provides us with the prospect of identifying a clear and fundamental area of disagreement between the theory theorist and the simulation theorist.

On this more relaxed version of the theory theory, the theory theorist insists that my ability to predict and explain what my conspecifics will think and do will depend, *inter alia*, upon a body of psychological knowledge. If, in accordance with the more relaxed view, the theory theorist is not going to put any great weight upon the precise way in which that body of knowledge is structured, then the theory theory expands to fill a larger region of logical space, leaving only a rather restricted area for alternatives. A genuine disagreement between the theory theory and an alternative view can be generated only if the alternative denies that our folk psychological ability depends upon our possession of a body of psychological knowledge, *simpliciter*. And this can seem like an untenable position for the imagined alternative since, surely, it cannot be denied that we do have psychological knowledge, and that, at least sometimes, it is deployed when we explain and predict other people's behaviour.

2.4 Folk psychology and folk physics

The apparent strength of the theory theory position can be seen if we consider the comparison that Stich and Nichols (1992) draw between folk psychology and folk physics. The idea that we have a folk physics is the idea that our ability to move and act successfully in the physical world — to negotiate, for example, interactions with objects that we come into contact with — depends upon our possession of a body of information about, *inter alia*, the ways in which physical bodies generally tend to behave. Thus, it has been argued in the psychological literature that adult reasoning about physical objects is informed by knowledge of principles about continuity and solidity. Our possession of this knowledge explains why it is that (Spelke et al., 1992, p. 607):

> . . . no subject in any study of physical reasoning has ever judged that any part of a material object would move discontinuously or would coincide in space and time with a second material object .

The claim being made by those who think of our ability to negotiate the physical world as being dependent upon knowledge of a folk physical theory is not just that our practical abilities can be given a theoretical description. Rather, the claim is that knowledge of the principles about continuity and solidity is causally implicated in our abilities. Our possession of this knowledge explains why it is that we have those abilities with respect to physical bodies.

Against the background of this comparison, the theory theorist of folk psychology says that, just as our ability to negotiate the physical world depends upon a body of knowledge about the ways in which physical objects tend to behave, so also our ability to negotiate the social world depends upon our possessing a body of knowledge about the way that people tend to behave. And now the dialectical position seems to be this. If the analogy with folk physics is a fair one, then any alternative to the theory theory of folk psychology must give us grounds for supposing that our understanding of the social world proceeds on a quite different basis from our understanding of the physical world.

3. The simulation theory

The theory theory involves a particular view of our epistemological relation to other people. According to that view, other people are objects in our environment, and the task of understanding them is no different, in principle, from the task of understanding the behaviour of other, more inert, objects (‘stars, clouds or geological formations’). This view provides a clear rationale for the comparison between folk psychology and folk physics.

The simulation alternative sets out from a different starting point; namely, the thought that when we try to understand other people we are trying to understand objects of the same kind as ourselves. This renders our epistemological situation when confronted with the behaviour of our fellows radically different from the situation when we are confronted with the behaviour of other objects. The similarity between the understander and the to-be-understood creates the possibility of a distinctive methodology (Heal, 1986, p. 137):

> I can harness all my complex theoretical knowledge about the world and my ability to imagine to yield an insight into other people *without any further elaborate theorising about them*. Only one simple assumption is needed: that they are like me in being thinkers, that they possess the same fundamental cognitive capacities and propensities that I do.

It might be possible to employ something like this methodology while still regarding the explanatory and predictive tasks as proceeding in terms of ‘how things generally tend to happen’. In that case, we would simply be using ourselves as measures of the way that things regularly occur: This is how things tend to happen with me; the other is much like me; so, this is how things will tend to happen with the other. But often, a simulation theory answer to our question (d) about the resources drawn upon in understanding other people is accompanied by a distinctive answer to question (c) about the key characteristics of our folk psychological practice.

According to this distinctive answer, what we are doing when we try to understand another person is not attempting to bring the other's behaviour under some law-like generalization, nor even to assure ourselves that there is some generalization that would subsume the events that confront us. Rather, we are trying to *make sense* of the other. Thus (Heal, 1986, p. 143):

> The difference between psychological explanation and explanation in the natural sciences is that in giving a psychological explanation we render the thought or behaviour of the other intelligible, we exhibit them as having some point, some reasons to be cited in their defence.

The simulation theorist's combined answer to questions (c) and (d) is then that the key characteristic of folk psychological practice is that it is a matter of rendering other people intelligible, and that the resources drawn upon include our knowledge of the world, and our ability to imagine, but not any body of distinctively psychological knowledge.

The role of imagining in this answer is important, since it cannot be taken for granted that simulation theory is the only way of developing the idea of folk psychological practice as involving a distinctive kind of explanation. The essential point is that the simulation theorist sees engagement in folk psychological practice as *re-enactment*.

Amongst philosophers working on the simulation approach to folk psychology, Robert Gordon (Gordon, 1986, 1992, 1995, this volume) has done as much as anyone to spell out just exactly how the simulation strategy works. One of his examples (Gordon, 1992) has me walking along a trail with a friend when that friend suddenly turns and runs in the opposite direction. My job is to understand this piece of behaviour. In accordance with the simulation strategy, I first imagine how the world looks from his position (the position he was in immediately before he turned and ran). In imagination (or perhaps even in reality) I move myself to that position. Suppose that when I do this I can see what looks like a bear moving towards us. What am I — that is, I identified in imagination with my friend — to do?

Perhaps the answer is that I decide to run. (So far as the project of understanding my friend's action is concerned, this decision is still in imagination, though I may also take my own decision to run away as well.) In that case, I can explain my friend's running away; and if I had made the imaginative identification a few moments earlier, then I might have predicted his running away.

But perhaps that is not the answer. Maybe the answer is that I decide to 'play dead' — I have learned from hunting manuals that this is the thing to do if confronted by a bear. In that case, I am not yet in a position to explain what happened; as yet, I have not made sense of my friend's action. I have to take on board relevant differences between my companion and myself, and then re-run the decision process. Although the first case — 'total' projection — is the default strategy for simulation, most often what is required is a 'patched' projection.

3.1 The first person case

The simulation strategy involves using imagination to cantilever out from our own theoretical and practical reasoning — leading to judgements and decisions — to an understanding of the beliefs and actions of another person. In imagination I go through some theoretical or practical reasoning and arrive at a judgement or a decision. If this is to yield a prediction about the beliefs or actions of another person then, of course, I need to be able to know about my own judgements and beliefs, decisions and intentions. But simulation theory itself does not, strictly speaking, provide an account of first person knowledge of mental states. Rather, simulation theory takes some such account for granted.

In fact, there are two rather different approaches to first person knowledge of mental states that have found favour with friends of simulation theory. Alvin Goldman (1989, 1993) develops the idea that mental states like belief have intrinsic, introspectible qualities — in short, qualia. In apparent contrast, Gordon (1995, this volume) sees first person judgements as involving an 'ascent routine' (this volume, p. xxx):

> [I]f someone were to ask me, (Q1) 'Do you believe Mickey Mouse has a tail?' I would ask myself, (Q2) 'Does Mickey Mouse have a tail?' . . . If the answer to Q2 is Yes, then the presumptive answer to Q1 . . . is Yes.

The issues between these two approaches are quite delicate. But, without entering upon a detailed comparison, we can make two remarks to suggest that the contrast might not be quite as stark as it initially appears. First, the idea that belief states have qualitative properties does not have to be coupled with the regressive idea of the subject having to match the properties of the belief with the properties of some mental 'colour card'. Second, engaging in an ascent routine clearly does not require possession of a concept of the mental state in question, if a concept is thought of as a mental template. But still, if an ascent routine is to yield a judgement to the effect that the subject believes that such-and-such, then the subject needs to possess the concept of belief that is deployed in that judgement.

Even if the idea that belief states have intrinsic, introspectible properties escapes a charge of regress, it is still apt to sound outlandish — as if belief states were being run together with sensations. But perhaps it is possible to make good sense of the idea by beginning from a thought about dispositional properties and their bases. A subject who (consciously) believes that, say, Mickey Mouse has a tail is in a position to judge that she has that belief. It is very natural to suppose that there is something about the belief state in virtue of which the subject is in a position to make that second-order judgement; and it is then quite natural to say that it is the belief's being a phenomenally conscious state that makes it accessible to the subject. What we do not need to say — what really would sound outlandish — is that the belief state's phenomenal properties are non-representational (or sensational) properties that are merely correlated with its representational (or semantic) properties. Rather, we should allow that representational properties can themselves be phenomenal properties.

This line of thought leaves unanswered important questions about the introspectible difference between different types of mental state with the same content — between believing and intending, for example. And there are important challenges for the idea of ascent routines, as well (see Carruthers, this volume). But, rather than pursue those questions and challenges, we turn to the point that the case of first person knowledge of mental states is liable to seem problematic for the theory theory too. Indeed, Goldman (1993) uses this problem for the theory theory approach to first person attribution in order to motivate his own preference for cognitive qualia.

The problem for the theory theory is that it does seem massively counterintuitive to suppose that first person attributions require checking that a mental state stands in a particular network of causal relations before pronouncing it to be a belief, say, that Mickey Mouse has a tail. In response to this worry, Carruthers (this volume) makes the important point that it is open to a theory theorist to adopt a perceptual — or more generally, non-inferential — account of our coming to make first person judgements, while still holding that our grasp upon the contents of the judgements arrived at is constituted by our knowledge of a theory in which the mental concepts figure.

With this variation in place, the theory theory is still offering two distinctive proposals. One proposal is an answer to our question (d) about the resources that are drawn upon in our folk psychological practice. Knowledge of a psychological theory is implicated in at least our third person attributions, explanations, and predictions. The other proposal is an answer to our question (a) about concept mastery. Our mastery of mental concepts — whether they are deployed in third person or in first person judgements — is constituted by knowledge of a psychological theory.

The simulation theory also offers a distinctive answer to the question about resources that are drawn upon in third personal folk psychological practice. The crucial ingredient is the capacity to engage in imaginative identification. But it is not so clear whether there is a distinctive simulation theoretic answer to the question about concept mastery.

3.2 Simulation and mental concepts

Goldman (1989) is not optimistic about the prospects of the simulation approach yielding an account of our mastery of mental concepts, but Gordon's more ambitious ('radical') version of the simulation approach does seem to be intended as an answer to a question about concept mastery. In his first paper, Gordon speaks of simulation theory as offering 'a way of interpreting ordinary discourse about beliefs' (1986, p. 166), but does not address the concept mastery question in anything like the terms that we have posed it. However, in his more recent work, he moves closer to these issues, and in his paper in this volume he offers the beginnings of an account of what simulation contributes to mastery of mental concepts.

Gordon makes use of a spatial analogy according to which a person's being in a mental state is to be thought of as the state's being '*at a mental location*' or '*mentally* at a location'. Thus (Gordon, this volume, p. xxx):

> When an ascent routine is used within the context of a simulation, a new logical space is opened. One can understand the object-level question . . . 'Does Mickey Mouse have a tail?' to have answers *at* various locations in this space. For example, one child, Jane, might simulate another, Mary, and then ask herself, in the role of Mary, the object level question, 'Does Mickey Mouse have a tail?' Simulation links the answer to the particular individual whose situation and behaviour constitute the evidence on which the simulation is based — the individual with whom one is identifying within the simulation. This, it seems to me, gives sense to the notion of something's being a fact *to* a particular individual.

Gordon's idea seems to be that our conception of Mary's believing that Mickey has a tail is a conception of the question 'Does Mickey have a tail?' having an affirmative answer at a particular point in a space, namely the point that Jane reaches in imagination by identifying with Mary. But the problem that is faced by any answer to the concept mastery question along these lines is that 'simulation is such a fallible procedure' (Goldman, 1989, p. 182).

Mary might not really believe that Mickey Mouse has a tail, even though Jane is indeed led to assert that Mickey Mouse has a tail within the scope of the best simulation of Mary that she can manage. In terms of the analogy with spatial location, the question might have an affirmative answer at the point that Jane reaches, but not at the point where Mary is really doxastically located. Because simulation is a fallible procedure, Jane may not reach the correct point.

If the simulation theorist is to make any further progress along these lines then, it seems, the analogy with spatial location needs to be coupled to a notion of idealised simulation. So, the revised proposal on behalf of

Gordon's radical simulation theory is this. Our conception of Mary's believing that Mickey has a tail is a conception of the question 'Does Mickey have a tail?' having an affirmative answer at the point that would be reached in an *ideal* simulation of Mary. But, as Peacocke (1994, p. xxv) points out, this notion of idealised simulation is fraught with difficulties:

> Now consider a supposedly idealized simulator of another's attitudes. The . . . threat is that, in so far as we can make sense of the idea, we have to draw on some prior understanding of what it is for another to have particular propositional attitudes.

So, the prospects for the most radical kind of simulation theory are still unclear.

4. The shape of the debate

If we leave the concept mastery question to one side, and just focus on the question about the resources that are drawn upon, then the debate seems to come down to this. The theory theory says that our ability to negotiate the social world depends upon our possessing a body of empirical knowledge about how people's situations, mental states, and behaviour are related. The simulation alternative needs to find a distinctive place for the ability to engage in imaginative identification while also denying that our folk psychological practice relies upon possession of a body of psychological knowledge.

There are then a number of ways in which the distinction between the two sides of this debate might be blurred. One particular kind of threat of collapse arises if the theory theory makes use of the idea of tacit knowledge of a theory. For the notion of possessing tacit knowledge has to be defined, and once a definition is offered it is a substantive question whether someone who engages in mental simulation — as that is described by Goldman or Gordon, say — counts as having tacit knowledge of a psychological theory. If the notion of tacit knowledge is defined so thinly that a simulator also counts as a tacit knower of a psychological theory then, in a quite strict sense, the distinction between the two sides collapses (see Davies, 1994; Heal, 1994; Perner, 1994, this volume).

But, even supposing that the basic distinction between simulating or re-enacting and drawing upon a body of psychological knowledge remains intact, still there are ways in which the theory theory approach and the simulation alternative might be argued to overlap.

4.1 Opposed or complementary approaches?

One line of argument that has been present from the outset of the debate is that a simulator needs to draw upon a body of psychological knowledge in order to carry through a simulation. Thus, Dennett (1981/1987, pp.

100–101):

> How can it [sc. simulation] work without being a kind of theorising in the end? For the state I put myself in is not belief but make-believe belief. If I make believe I am a suspension bridge and wonder what I will do when the wind blows, what 'comes to me' in my make-believe state depends on how sophisticated my knowledge is of the physics and engineering of suspension bridges. Why should my making believe I have your beliefs be any different? In both cases, knowledge of the imitated object is needed to drive the make-believe 'simulation' and the knowledge must be organized into something rather like a theory.

Consider again Gordon's example in which my friend and I are walking along the trail. Surely, the theory theorist will say, in order to explain or predict my friend's action of running away, I need to know something about the typical causal relations between recognising a bear, being afraid, and taking evasive action. In addition, I need to know something about my friend's psychological make-up, and that knowledge, too, will be dependent upon pieces of theory (about the attitude towards bears that tends to be produced by a certain kind of education, for example).

The simulation theorist has a number of responses to make at this point. Particularly, the simulation theorist argues that the theory theorist is making an unwarranted and unparsimonious assumption. This assumption is that, over and above any actual thinking (say, thinking about tracks, bears, and escape) that takes place — whether *in propria persona* or within the scope of a simulation — there are also general psychological principles, knowledge of which explains the movements of thought that occur during a period of thinking, or within some episode of simulation. The simulation alternative sees no need for these two layers of thought: a layer that is the actual episode of thinking about the world, and a layer of meta-thinking that brings about the movement of thought in the first layer. All that is needed, according to the simulation theorist, is that some thinking take place, in accordance with the canons of rational cognition. The dynamics of thought require no meta-cognitive engine.

This line of argument is used to reject the suggestion that mental simulation will inevitably be 'theory driven' rather than 'process driven' (Goldman, 1989). But it can also be used to undermine the theory theory itself. For what the theory theory appears to be committed to is not just knowledge of some general psychological principles, but also knowledge of indefinitely many indefinitely detailed principles about thought concerning specific subject matters (Heal, 1995, this volume). Botterill (this volume), Carruthers (this volume), and Perner (this volume) all concede, on behalf of the theory theory, that the intrusion of something like simulation will be needed — what Carruthers calls 'simulation within a theory' and Perner calls 'content simulation'.

On the other side, advocates of the simulation theory typically acknowledge that inductively based generalizations play a role in real life use of mental simulation (Goldman, 1989; Harris, 1992), and Perner (this volume) presents empirical data that points in the same direction. So, to that extent, we seem to be moving towards a measure of agreement over the need for hybrid theories. Future research will need to address in detail the ways in which simulation and deployment of knowledge interact (Perner, this volume, p. xxx):

> [S]ince any theory use involves an element of simulation and since simulation on its own cannot account for the data, the future must lie in a mixture of simulation and theory use. However, what this mixture is and how it operates must first be specified in some detail before any testable empirical predictions can be derived.

4.2 Cognitive penetrability: The crucial test?

The point about the relative lack of economy involved in adoption of the theory theory approach is made vivid by an example introduced by Paul Harris (1992). Suppose we are asked to predict the grammaticality judgements that a speaker of the same language would make when confronted with a range of sentences in the language. Harris reasonably claims that predictive success would be high, and offers this explanation (1992, p. 124):

> 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. The proposal that you have two distinct tacit representations of English grammar, a first-order representation that you deploy when making your own judgements, and a metarepresentation (i.e. a representation of other people's representations) that you deploy in predicting the judgements made by others, so designed as to yield equivalent judgements, strains both credulity and parsimony.

Furthermore, we might suppose, what goes for the prediction of grammaticality judgements goes also for the prediction of belief formation on the basis of inference, and for the prediction of intentions and behaviour.

However, Stich and Nichols (1995) respond to Harris's example by distinguishing cases. They are inclined to make concessions to the simulation theory over the explanation and prediction of a person's judgements and beliefs in a particular perceptual situation, and perhaps also in the case of belief formation on the basis of inference. But they argue strongly for the theory theory as providing a better account of the prediction of behaviour. Given the role of theoretical inference in any decision taking process, we are not convinced that different stances towards prediction of belief formation on the basis of inference and prediction of behaviour can be justified. But that is not our main concern here.

Stich and Nichols' argument hinges on the phenomenon of cognitive penetrability, introduced in their earlier paper (1992). The issue is this. If predictions are based upon deployment of a theory (a body of information or misinformation), then those predictions are liable to be false if the theory is incorrect in any way. Theory based predictions are subject to error introduced by misinformation. But, a flawed theory will obviously have no impact upon predictions that do not draw upon it. Even if I have deeply flawed psychological views still, if I use mental simulation to generate predictions about what people will do given their beliefs and desires, then my flawed theory need introduce no error into my predictions.

So the crucial empirical question is, apparently, whether we are liable to make false predictions about other people's decisions, intentions, and actions. If we are then, according to Stich and Nichols (1992, 1995), this favours the theory theory. And, as they point out, there are indeed examples where folk psychological prediction lets us down.

One of the examples used by Stich and Nichols (1992) involves predictions about what subjects will do when invited to select amongst items all of which are, unknown to them, identical in quality. Another example concerns predictions about the price for which subjects will sell back raffle tickets that were obtained in two different conditions. In both cases, predictors consistently give incorrect answers.

Goldman (1992) and Harris (1992) point out responses that are available to the defender of the simulation theory. The general idea of these responses is this. The simulation method will only arrive at correct answers if it begins from the correct inputs, and the inputs that the predictors use are very largely determined by the way that information about the original subjects' situation is presented to the predictors. Nichols, Stich, Leslie and Klein (this volume) report a more formal experiment, designed so as not to be open to the objections that Goldman and Harris raise. In this more tightly controlled setting, predictors still make errors.

However, there is a possible source of error in predictions that is still not addressed by these experimental findings. For, as Harris points out (1992, p. 134) there may be purely mechanical influences on decision taking that are not captured by mental simulation. Indeed, we can see how these influences might be relevant, even in the sort of case where Stich and Nichols (1995; and Nichols et al., this volume) are inclined to make concessions to the simulation theory.

We are able to use our own perceptual abilities and our own inferential abilities in simulation mode in order to arrive at predictions about another's belief formation. But, if we learn that the other has just ingested a certain substance, then we cannot make allowances for that fact by imaginative identification alone. Unless we are prepared to ingest the same substance ourselves, we need to be told what effect it has upon the processes of perception and inference. What is needed, then, is a little piece of theory — some information about non-rational influences upon psychological processes.

Exactly the same goes for the processes of decision taking. Unless we are ready and able to subject ourselves to the same non-rational influences that affect those whose decisions we seek to predict, we shall need to augment our ability to simulate with some pieces of theory. This has always been accepted by the friends of mental simulation. As Heal says (1986, p. 139):

> Replication [simulation] theory must allow somewhere for the idea of different personalities, for different styles of thinking and for non-rational influences on thinking.

What the simulation theorist denies is that we need to draw upon an empirical theory about rational processes of belief formation or decision taking.

Conclusion

The mental simulation debate has reached a stage at which there is considerable agreement about the need to develop hybrid theories — theories that postulate both theory and simulation, and then spell out the way in which those two components interact. But the situation is complicated by the fact that there are several quite different versions of the theory-theory and of the simulation alternative. On the side of the theory-theory, we have already seen (Section 2 above) that there is a range of options, and certainly there are important differences between the approaches taken by, for example, Perner (1991), Wellman (1990) and Leslie (1987). On the side of the simulation theory, Gordon (1986), Heal (1986) and Goldman (1989) each stress rather different ideas. Selective borrowing from these accounts would allow the development of a huge variety of hybrid theories.

We would like to draw attention to a type of theory that, on the one hand, does not belong in the theory-theory camp but, on the other hand, leaves out the key idea of at least Gordon's version of the simulation theory. This type of theory stresses both the importance of first order thought about the world and the idea of a distinctive kind of explanation — making sense of another person — but does not regard either explanation or prediction as essentially involving imaginative identification with the other.

When I am considering how to act in a given situation, I bring to bear my knowledge about the world, and arrive at a judgement about what is *the* *thing to do*. Both knowledge and imagination are certainly drawn upon, but there need be no intrusion into my own decision taking of any body of empirical theory about psychology — about what people in certain situations and with certain propositional attitudes generally tend to do.

Then, when I turn to the task of explaining or understanding the actions of another person, just the same kind of normative judgement is relevant (McDowell, 1985, p. 389):

> [T]he concepts of the propositional attitudes have their proper home in explanations of a special sort: explanations in which things are made intelligible by being revealed to be, or to approximate to being, as they rationally ought to be.

Once again, the imagination is liable to be involved. But there is no special role for imaginative identification in this transition from the first to the third person, since the normative judgement about what is the thing to do is not an essentially first person judgement.

The task of predicting the behaviour of another does require something more than just this kind of normative judgement. It requires the assumption — which might well operate as a kind of default — that the other will indeed do the thing that is the thing to do (Heal, 1986, p. 137):

> Only one simple assumption is needed: that they are like me in being thinkers, that they possess the same fundamental cognitive capacities and propensities that I do.

But there is still no special role for imaginative identification. So, this type of theory differs from at least Gordon's version of the simulation theory. It is also different, surely, from the theory-theory. The single empirical assumption that people will, on the whole, do the sensible thing scarcely amounts to the body of psychological knowledge that theory-theorists envisage.

There may well be some kind of theory involved in the process of arriving at a judgement as to what is the thing to do. Whether there is or not is not an obvious matter. But if there is, then it is not an empirical psychological theory about what people generally tend to do. As Simon Blackburn says (1992, p. 195):

> In one way or another the fact that we need to theorise under a 'principle of rationality', or to see a proper point in people's doings in order to understand them, marks off this kind of theorising from anything found in the natural sciences.

We cannot now offer any evaluation of this alternative position, nor explore its extension to the explanation and prediction of another person's beliefs. We are not even sure whether it should be regarded as genuinely distinct from both the theory-theory and the simulation theory, or rather as a variation on the simulation theory theme. (It would appear to count as a version of the simulation theory on Stich and Nichols's (1992, p. 47) way of 'drawing the battle lines', though they assume that there are only two alternatives to be considered.) But we do think that it deserves consideration.

References

The papers marked * are reprinted in:

Davies, M. and Stone, T. (eds) 1995: *Folk Psychology: The Theory of Mind Debate*. Oxford: Blackwell Publishers.

*Blackburn, S. 1992: Theory, observation and drama. *Mind and Language*, 7, 187–203.

Churchland, P.M. 1988: *Matter and Consciousness* (Revised Edition). Cambridge, MA: MIT Press.

Collingwood, R.G. 1946: *The Idea of History*. Oxford: Oxford University Press.

Davies, M. 1994: The mental simulation debate. In C. Peacocke (ed.), *Objectivity, Simulation and the Unity of Consciousness*. *Proceedings of the British Academy*, 83, 99–127.

Dennett, D.C. 1981: Making sense of ourselves. *Philosophical Topics*, 12, Number 1; reprinted as J.I. Biro and R.W. Shahan (eds), *Mind, Brain, and Function: Essays in the Philosophy of Mind*, Brighton: Harvester Press, 1982, 63–81. Reprinted in *The Intentional Stance*. Cambridge, MA: MIT Press, 83–101.

Fodor, J. 1987: *Psychosemantics*. Cambridge, MA: MIT Press.

*Goldman, A.I. 1989: Interpretation psychologized. *Mind and Language*, 4, 161–85.

Goldman, A.I. 1993: The psychology of folk psychology. *Behavioral and Brain Sciences*, 16, 15–28.

*Gordon, R.M. 1986: Folk psychology as simulation. *Mind and Language*, 1, 158–71.

*Gordon, R.M. 1992: The simulation theory: Objections and misconceptions. *Mind and Language*, 7, 11–34.

Gordon, R.M. 1995: Simulation without introspection or inference from me to you. In T. Stone and M. Davies (eds), *Mental Simulation: Evaluations and Applications*. Oxford: Blackwell Publishers, 53–67.

*Harris, P.L. 1992: From simulation to folk psychology: The case for development. *Mind and Language*, 7, 120–44.

*Heal, J. 1986: Replication and functionalism. In J. Butterfield (ed.), *Language, Mind and Logic*. Cambridge: Cambridge University Press, 135–50.

Heal, J. 1994: Simulation vs. theory theory: What is at issue? In

C. Peacocke (ed.), *Objectivity, Simulation and the Unity of Consciousness*. *Proceedings of the British Academy*, 83, 129–44.

Heal, J. 1995: How to think about thinking. In T. Stone and M. Davies (eds), *Mental Simulation: Evaluations and Applications*. Oxford: Blackwell Publishers, 33–52.

Heider, F. 1958: *The Psychology of Interpersonal Relations*. New York: Wiley.

Leslie, A.M. 1987: Pretense and representation: The origins of 'theory of mind'. *Psychological Review*, 94, 412–26.

Lewis, D. 1972: Psychophysical and theoretical identifications. *Australasian Journal of Philosophy*, 50, 249–58. Reprinted in N. Block (ed.), *Readings in Philosophy of Psychology, Volume 1*. London: Methuen, 1980.

McDowell, J. 1985: Functionalism and anomalous monism. In E. LePore and B.P. McLaughlin (eds), *Actions and Events: Perspectives on the Philosophy of Donald Davidson*. Oxford: Basil Blackwell, 387–98.

Peacocke, P, 1994: Introduction: The issues and their further development. In C. Peacocke (ed.), *Objectivity, Simulation and the Unity of Consciousness*. *Proceedings of the British Academy*, 83, xi–xxvi.

Perner, J, 1991: *Understanding the Representational Mind*. Cambridge, MA: MIT Press.

Perner, J. 1994: The necessity and impossibility of simulation. In C. Peacocke (ed.), *Objectivity, Simulation and the Unity of Consciousness*. *Proceedings of the British Academy*, 83, 145–54.

Pettit, P. 1986: Broad-minded explanation and psychology. In P. Pettit and J. McDowell (eds), *Subject, Thought, and Context*. Oxford: Oxford University Press, 17–58.

Spelke, E.S., Breinlinger, K., Macomber, J. and Jacobson, K. 1992: Origins of knowledge. *Psychological Review*, 99, 605–632.

Stich, S.P. 1981: Dennett on intentional systems. *Philosophical Topics*, 12, Number 1; reprinted as J.I. Biro and R.W. Shahan (eds), *Mind, Brain, and Function: Essays in the Philosophy of Mind*, Brighton: Harvester Press, 1982, 39–62.

*Stich, S.P. and Nichols, S. 1992: Folk psychology: Simulation or tacit theory? *Mind and Language*, 7, 35–71.

Stich, S.P. and Nichols, S. 1995: Second thoughts on simulation. In T. Stone and M. Davies (eds), *Mental Simulation: Evaluations and Applications*. Oxford: Blackwell Publishers, 87–108.

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

Wolpert, L. 1992: *The Unnatural Nature of Science*. London: Faber.