Proceedings of the British Academy, **83**, 129–144

COMMENTARY

Simulation vs. Theory Theory: What is at Issue?

JANE HEAL
St. John's College, Cambridge, CB2 1TP

IN COMMENTING ON Martin Davies' paper I would like to take up and expand on one of the issues he raises, namely the possibility that the dispute between simulationists and theory theorists should turn out to be illusory. I agree with Davies that this possibility does indeed exist, but I shall argue that it is more difficult than he suggests to avoid it becoming actual.[1] In order to see that there is indeed a dispute, and what might really be at issue in it, we need to pay close attention to how it is conceptualized. If we can get clearer on this, then it may be easier to make progress on the other fascinating and central topics Davies mentions, for example whether it is possible to combine simulationism with a proper recognition of the first/third person asymmetry in criteria for psychological ascriptions, and whether simulationism has anything to offer on the nature of mental states or on the possession conditions for psychological concepts.

Let us start by reminding ourselves of what it is to have tacit knowledge of some theory, at least of what is said in the best account that we currently have of the matter, namely that offered by Davies himself (1987), in development of a suggestion of Evans. This account says that a person can be credited with tacit knowledge of a theory

Read at the British Academy 13 March 1993.

[1] I would like to acknowledge that I was stimulated to the following thoughts by D. H. Mellor's resolute refusal to concede that there could be an important difference between theory and simulation.

consisting of a set of propositions provided that we have empirical evidence that there is, for each separate proposition, a corresponding separate element inside the person which mediates causally between explicit premises and explicit conclusions; it is further required that the overall causal pattern in this structure duplicate the logical pattern of the relations within the theory. The empirical evidence for tacit knowledge will thus come in the form of actual and counterfactual patterns among observable manifestations, namely those explicit beliefs which, on the tacit knowledge hypothesis, are inferred with the aid of the tacit theory. For example, it might be the case that if a person loses the commitment to one explicit belief which is (on the hypothesis under consideration) derived with the help of some particular proposition of the tacit theory, then he or she also at the same time loses commitment to all the other explicit beliefs which depend upon that same proposition as premise. Or it might be the case that if a person changes an explicit belief which supposedly depends on one tacit premise then all the beliefs similarly dependent change in a correlated way. To put matters in a nutshell, if the logical structure of the supposed tacitly known theory (in the form of presupposition, exclusion, implication, etc.) is paralleled by an isomorphic and well articulated causal structure, then that is necessary and sufficient for attribution of tacit knowledge.

Davies sketches a line of argument (in his paper in this volume) which suggests that it is important to be careful how we describe simulation, on pain of having the supposed difference between the simulation and the theory theory account of psychological understanding collapse on us. Let us concentrate upon the case where one person, A, is trying to predict the future thought or action of another, B, on the basis of information about B's current psychological state. (There are many other cognitive tasks having to do with understanding others' psychological states, for example, arriving at judgements about them on the basis of behaviour, retrodicting them, explaining them, etc. But the matter of prediction is one which all simulationists agree in thinking to be a strong case for them and one which, they say, their view and theory theory would handle differently. It is thus a good focus for our discussion.) The theory theorist says that A's prediction is produced by A's application of his tacit theory about psychological states, how they interact and what they give rise to. The simulationist, by contrast, says that A simulates B's initial state and then allows some process to unroll in him which ends up with him having some simulation of B's

future intention. This unrolling is supposed to be quite different from working out the implications of a (tacit) theory; rather it is said to be driven by the same processes which carry forward the actual thinking which we do on our own behalf.

Suppose now, says Davies, that we fill in the detail of the simulationist account by saying that for A to simulate 'B believes that *p*' is for A to imagine 'I believe that *p*'. And suppose further that this mental state (together with simulations of B's other states) unrolls into some representation of an intention — 'I intend to *V*'. Now it may be that the derivational process is isomorphic to the actual thinking as it unrolls in B. But, as Davies implies, this alone is not enough to vindicate a non-theory approach. After all, as Goldman (1989) originally pointed out, a good *explicit* theory enables us to produce an unfolding sequence of representations which runs parallel to developments in the item to be understood. What is crucial is the nature of the processes underlying the derivations. But all we have said so far about the imagined 'simulation' leaves it open that there exists some systematic relations of dependence between input and output (i.e. between what A starts out thinking about B and what he ends up thinking about B) which mirror in their causal structure the logical structure of a psychological theory. But, if this is so, then, on the above account of tacit knowledge, it will be reasonable to postulate that A is calling upon a tacit theory. The fact that phenomenologically A is inclined to say that he is empathizing with B or recreating B's thoughts or some such, is neither here nor there.

As Davies remarks, it is not inevitable that simulation and theory theory turn out indistinguishable, if we take 'simulation' in the way outlined. But the danger is particularly great because the input and output to the supposed simulation process are both explicitly representations of people having psychological states. (They are imaginings with the contents 'I believe that *p*', 'I intend to *V*', etc.) Given this, then it seems likely that we shall discover certain patterns of causal dependence between input representations and output representations. And it is also probable that the pattern will have an overall shape which strongly suggests interior mediating structures of the kind which in turn license attribution of knowledge of a tacit theory. This is so because, *ex hypothesi*, we are imagining that the predictions in question are mainly successful and we are also imagining that there could be some theory which would produce the same, i.e. the successful, predictions. So (unless we start putting extra conditions on what is required

for theory possession) it looks as though the basic data which the simulationists and theory theorists both acknowledge, namely the structured relations of dependence between explicit input and explicit output, are going to support the theory theorist's claim. Thus the simulationist's idea that he had articulated some distinctively different position begins to seem problematic.

But, Davies suggests, we can stave off the threat of collapse by insisting that when A simulates 'B believes that *p*' we take it that A's imaginative enterprise consists of imagining that p and *not* imagining 'I believe that *p*'. If we thus remove the content 'I believe' from the content of what is imagined then collapse will not occur because (says Davies) 'processing mechanisms that mediate transitions amongst states with such contents are not going to be embodiments of tacit knowledge of the principles of a psychological theory.'

Now I agree with Davies that, in the case set out above, there is danger of collapse. But I shall try to argue that it is not so much one or the other exact description of what simulation is (what its content is) which introduces the danger, but rather something lying unarticulated behind this, namely a particular but non-obligatory conception of the nature of the whole question. Davies alludes briefly to this in contrasting Goldman's willingness to talk of 'off line' use of psychological mechanisms with Gordon's preference for the 'unscientized' notion of imaginative identification. I shall explore one thing which might be meant by this contrast, in so doing aligning myself with Gordon as against Goldman (although I am not sure that Gordon understands the contrast quite as I do or would adduce the reasons which I shall bring forward).

There are, I suggest, two ways of looking at the theory theory vs. simulation debate. On one approach it is seen as an empirical question about how our undoubted ability to predict others' future thoughts, feelings and actions on the basis of knowledge of their current psychological states is implemented at a sub-personal level. On the other approach it is seen as a question about how abilities or capacities at a personal level are interrelated. This second idea will, for the moment, seem obscure. I shall endeavour to clarify it later. But first I want to try to make plausible the idea that the threat of collapse is induced not so much by ways of specifying what it is to simulate as by the conception of the question as being empirical and about sub-personal mechanisms.

Clearly objects — stars, atoms, bank rates and babies — are one

thing and thoughts about them another and very different thing. Stars, atoms, etc. are out in the world behaving in their own complicated and distinctive way, while thoughts are in people's heads (or minds) behaving in other very different ways and making these people speak, move, etc. So it looks as if thinking about objects and thinking about thoughts could involve quite separate parcels of knowledge, as separate as the parcels involved in thinking about bank rates and thinking about babies. Of course all parcels of knowledge will require some basic notions like 'thing', 'property', and 'time' in common. But apart from this core of categories and the principles which go with them, the parcels may be entirely distinct. Knowing about bank rates will not help me in dealing with babies and equally (so this line of thought runs) knowing about bank rates will not help me in dealing with your thoughts about bank rates. Rather it is in virtue of possessing a special body of knowledge about psychological matters that we are able to predict others' thoughts and behaviour.

This is the picture behind the theory theory. The simulationist, on the first approach to the issue, does not have any real quarrel with the broad outlines of the picture. He allows that objects and thoughts are extremely different and that it is logically possible that we proceed by having a special theory of the latter. But, he says, our facility in predicting others might be explained a different way. We ourselves have minds, in which thoughts occur. Now if it were possible to 'unhook' our minds (or part of them, say the Practical Reasoning System) and use them 'off line' then we could derive predictions about others without use of special theory. All we would need to do is note the others' thoughts, feed them in to our unhooked mind, note what it comes out with and attribute it to the other. We do not need to have any separate parcel of knowledge about minds to make this work; we just have to have minds which we can use in a certain, somewhat non-standard, way.

Note that on this kind of story it is (logically at least) possible that we should use the same strategy for other non-mental items, provided we have specimens of them inside us. So we cannot use anything like this 'simulation' method to understand galaxies but we might use it for hearts. If I could unhook my heart, feed it with pretend inputs and observe its output, then I could use it in a non-standard way to facilitate prediction of others' cardiac behaviour. (The assumption must be, of course, that my heart and their hearts are relevantly similar if the process is to produce accurate answers. For the sake of the example

we shall allow ourselves this background condition, together with a good deal of further physiological implausibility.) We could elaborate the fantasy by imagining that the whole process is carried on by some convenient subconscious mechanisms. So I am not aware of the temporary unhooking, the devising of suitable inputs or the recording of the output. All that happens, as far as I can tell, is that I possess some information about the other's situation (e.g. that he has climbed three flights of stairs) and pose myself the question what will happen to his or her heart; then I am (perhaps) vaguely aware of some physiological perturbations in me, after which the desired answer ('heart rate accelerates to 120 per minute') pops up in my mind. (All this sounds pretty risky; perhaps it would be better to carry round a spare heart to do the experiments on.)

But if this is how we think mental simulation works to deliver answers to psychological questions about others, is there really any significant difference between it and the theory theory? I suggest that there is none, at least on the definition of tacit knowledge we have assumed. Considering the heart case again, the spare or unhooked heart on which the experiments are done will have various independent features (size, layout, muscle resilience, etc.) each of which contributes in its own way to the determination of the output for a given input. In other words, the heart has a structure and mode of working which a good theory will record in its sentence by sentence specification. My verdicts on heart behaviour are systematically and counterfactually dependent on the interlocked working of these various features. If my heart were different in one of the features recorded by the imagined explicit theory then a whole class of verdicts would come out differently, while others not dependent on that feature would remain unchanged. The verdicts thus show just the same patternedness (e.g. of standing, falling and varying together in overlapping groups) as if they were delivered by a theory, the distinct axioms of which recorded the separate structurally important features of the heart. So there is an element in me playing a causal role analogous to the logical role of each statement of such a theory, namely the actual feature of the heart which does the mediating. Each such feature could change, independently of the others, for example if I have an illness or operation. And if it does then systematic changes in verdicts — exactly parallel to the changes which would occur were I to alter an axiom of the explicit theory — then follow. My suggestion is that I shall count as having a tacit theory of the heart in virtue of possessing a heart which I can

interrogate. The better discriminated my questions and answers to it, the better the theory it embodies for me. And because I use it to derive (correct) information about hearts other than mine and because its structures thus mediate whole collections of particular predictions, it counts as an encoding of generalizations about hearts.

We must acknowledge that this case of tacit knowledge is interestingly different from the classic cases (e.g. tacit knowledge of semantic theory enabling me to understand new sentences) in at least two ways.[2] One is that a tacit theory embodied in an inner model seems guaranteed to be true (at least if we allow the similarity of the inner item to the others in the class it is used to predict). Another is that we seem to make no sense of the idea that I could forget part of my tacit theory. But these points seems to suggest that the kind of tacit knowledge possessed in virtue of having an interior item on which to experiment is superior to the ordinary kind. It does not seem to show the impropriety of talking of tacit theory.

On reflection we can see why we have arrived at this upshot, namely of the collapse of our supposed dispute. What could carry or encode more information about a type of object than an object of that type itself? And that information is 'present' or 'available' to a person if he or she is able to extract it easily. The whole point about speaking of tacit theories is to stand back from commitment to explicit knowledge and also from commitment to the forms in which information is carried. Anything which fills the right logico-causal role is to count as a vehicle of the (tacit) knowledge. Thus the object itself must do so, if, as imagined, we carry it round inside us and can in fact interrogate it effectively.

Getting this unfortunate result (unfortunate at least if we think that the simulation/theory dispute as genuine and important) does not depend upon what we take to be the contents of imagination when one person simulates another. In the argument above I have said nothing at all about what it is like phenomenologically to do the unhooking and simulating. For all we have seen, simulating another believing that *p* could take the form of imagining that *p*. What is the difficulty then with the solution to the problem of collapse suggested by Davies? He thinks that once we have insisted that it is 'that p' which is represented in the simulator, rather than 'I believe that *p*', then we do not have any risk that he or she will turn out to have a tacit

[2] I am grateful to Christopher Peacocke for drawing these points to my attention.

theory; we only have the risk, argues Davies, if we start with a representation of the form 'someone believes that *p*'. But the trouble with the move is this. Unless I lose grip on the distinction between myself and others, I do start, and must start, with a representation having the content 'So and so believes that *p*'. My subsequently imagining that *p* (if that is what I do when I simulate) is only part of a total thought state which remains a thought about the other's thought. And what is delivered out at the far end of my deliberations is likewise an explicit representation of the other's future thought or action. This is the basic fact we are to explain, namely our facility in psychological prediction of others on psychological premises. And it is the nature of the patterned dependencies between input and output, and the interior causal structure those patterns lead us to postulate, which will justify attributions of tacit knowledge. If the theory happens to be encoded in us in an unhooked mind (or a spare mind, carried round for purposes of predicting others), so much the more ingenious.

On this whole way of conceiving the matter (i.e. as empirical and about sub-personal mechanisms) we shall have to redefine the notion of 'tacit theory', in order for there to be a question at issue. Stich and Nichols wish to make it turn on whether 'prediction, explanation and interpretation are subserved by a tacit theory *stored somewhere other than in the Practical Reasoning System*' (1992, p. 47, n. 7). If it is, then the theory theory wins; if not, not — according to them. So, on their way of looking at the matter, there is an important difference between the case where I predict by unhooking bits of my one and only mind and the case where I do it by having a spare mind to experiment on. In the latter case I count as having a theory, in the former not. It seems fairly clear that Stich and Nichols are committed to this since they emphasize that a theory can be a 'non-sentence-like, non-rule-based module which stores the information that is essential to folk psychological prediction and explanation' (1992, p. 47, n. 7), and a spare mind looks like just such a module. But although on this recasting of matters we would have a dispute, it is not clear how anything very central to the philosophy of mind could hang on it.

Another (and I think more promising) move by which a genuine dispute could be re-introduced would be to insist that a 'theory' should have a sentence-like mode of representation. There are undoubtedly many genuine and fascinating questions about vehicles of representation, the contrasts between analogue and digital, the potential of pictures, models and diagrams as opposed to sentences, how to define

these contrasts, etc. But if we take the theory vs. simulation debate to be bound up with these issues then it is resoluble only if two conditions are fulfilled:

1 we can make clear what the requirements are for a representation to count as a theory rather than a simulation and

2 brain or cognitive architecture turns out to be such that it actually delivers an unambiguous answer to the question of which is occurring.

My suspicion is that there will be an enormous variety of possible views on (**1**), and difficulty in motiving (from the point of view of philosophy of mind) interest in one rather than another, while on (**2**) the brain may not be very accommodating in conforming clearly to one or other of the a priori distinctions we devise under (**1**). (What if it all turns out to be a great connectionist tangle inside our skulls?) In short, on this whole way of laying out the dispute there is considerable risk that it will run away into the ground, with dispute definitions and difficulty of empirical resolution. I do not say that this upshot is inevitable. There are empirical studies which suggest the possibility of resolving some kinds of questions about ways in which information is encoded in the brain.[3] And perhaps studies of such things as the relative difficulty of various kinds of predictive tasks, the sorts of mistakes made and the like, will prove fruitful.

There would however still be considerable attraction in another conceptualization of the question which did not run this risk of having the dispute dissolve away into various different and possibly rather parochial issues. It seems to me also that another conceptualization might well capture better the issues that at least some of those who have written on the matter wished to raise. Let me try to sketch such an approach. The previous line of thought started with both disputants agreeing to the idea that babies and bank rates were one thing and thoughts about babies and bank rates quite another and agreeing also that it was possible that there should be a theory of thoughts about babies which was quite separate from any theory about babies. But what if the simulationist were to disagree with the theory theorist already at this point? He or she might maintain that it is impossible to separate thinking about thoughts from thinking about their objects in the way envisaged. Rather, it would be said, the capacity to think about thoughts is (and must be) an extension of the ability to

[3] For example, Johnson-Laird's work on mental models (1983).

think about their objects. Thus in thinking about someone else's thought that *p*, I must (in the central and usual case at least) exercise the same cognitive skills that I exercise when I myself judge that *p*, wonder whether *p* and the like.

Two features are characteristic of this approach to the dispute. One is that the matter is seen as well towards the a priori end of the spectrum of questions; the other is that the debate is seen as having to do with how we divide up and individuate the capacities which people have as people. In brief, the debate is about the structure of our conceptual scheme, in so far as it has to do with persons and their abilities.[4]

I should make clear here that the claim I am advancing here is, in the first instance, about what goes on when ordinary people in ordinary life think about the thoughts of others. So it is a claim about what it is to 'have the concept of thinking that *p*' where this just means the ability to understand what is said when another is said to think that *p*, the ability to see what difference this may make to that person's other thoughts and behaviour and so forth. It is a further question whether any of us do, or in the future will be able to, think about others' thoughts in some different way. This extra question has to do with what thoughts are like 'in themselves' so to speak and whether there might be some theoretical account of that. But the current discussion does not bear directly on that issue. The suggestion is rather that, in ordinary life, to learn to think about others' thoughts is to learn to apply, in a special way, ones own ability to think about the subject matter of those thoughts.

Two further examples may help to make clearer the nature of the idea of one capacity being identical with or an extension of another. Consider first belief and desire. If someone desires that *p* then he exercises a capacity, viz. to have that desire; and if he believes that *p* then he exercises a capacity to believe that *p*. But what is the relation between these two capacities? Can we envisage them as being separate

[4] It is not an implication of this that empirical studies, e.g. of development in children or of adults' success in tackling various cognitive tasks, is irrelevant to the simulation vs. theory dispute. For one thing, to present something as potentially recommendable by a priori considerations is a different thing from actually demonstrating it and I do not here claim to do this latter. Empirical studies may confirm the truth of a hypothesis and so encourage the search for a fuller proof. Empirical studies may also have a role in suggesting details of the elaboration of the proposal. And finally empirical data are always relevant to the whole question of the viability of the conceptual scheme itself. I do not wish to be committed to some sharp analytic/synthetic distinction.

in the way that, for example, the capacity to ride a bicycle and the capacity to extract square roots are separate? Clearly a person may be able to do one of these without the other and vice versa. Is it the same with believing and desiring? Surely it is not, because the two notions arrive together as inseparable parts of a unified package in terms of which the, in some sense prior and more fundamental, notion of purposive action is to be explained. One very fundamental connection is that the onset of the belief that *p* is what brings to an end action directed at bringing it about that *p*. So desiring that *p* and believing that *p* are better conceived as alternative manifestations of one underlying capacity, namely that of thinking about (representing) the state of affairs that *p*. Let us note also that if believing that *p* involves some intellectual sophistication, for example, the state of affairs that *p* is recognized as complex, then equally desiring that *p* has analogous sophistication; for example, if an agent comes to think that she has secured part of what is required for it to be the case that *p* then her desire that *p* will guide her actions towards securing the residual elements.

We may concede that a person might be capable of some particular desire without being capable (perhaps for Freudian reasons) of ever recognizing that the desire is satisfied. But this sort of case is necessarily abnormal. If we try to imagine what it would be like for it to be general we find we have imagined away all intelligently directed and effective action and so imagined away any possible subjects of either belief or desire.

Another case of interrelated capacities, in considering which we move importantly closer to the issue of simulation, is that of thought about the merely possible and thought about the actual. I mean 'thought about the merely possible' in a rather basic way here. It does not necessarily involve explicit employment of the concept of possibility. Rather it occurs (in the sense I am interested in) whenever a propositional content appears in a judgement without itself being affirmed. So it occurs when someone thinks 'If I *X* then *q*, but if I do not *X* then *r*' or 'It seems that *p* but is it really so?'. Thus thought about the possible is bound up with both our awareness of ourselves as creatures faced with decisions about the future, and also with our awareness of our epistemic limitations.

We can, perhaps, imagine creatures who are capable of representing the actual, inasmuch as they respond appropriately (by advance, retreat, etc.) to the things that confront them, but who are not capable

of deliberation about the future (i.e. representation of the various possibilities open to them) or of questioning their own judgements. So it would not be right here to say, as we did for belief and desire, that thought about the possible and thought about the actual are alternative manifestations of the one underlying capacity. But it is not plausible either to say that ability to think about the merely possible (i.e. to entertain unaffirmed contents) is a quite separate matter from ability to think about the actual (i.e. to make straightforward affirmative judgements). It makes no sense to suppose that a person should be able to think about the possible but not about the actual. (There are probably many good arguments for this view. Here is one very brief one. A person cannot act without having some non-conditional beliefs, i.e. some beliefs about the actual. So a putative person with only thoughts about the possible and none about the actual would turn out to be a non-agent and so a non-person.) Our conclusion must be that ability to think about the possible presupposes an ability to think about the actual and hence is an extension of it; someone thinking about the possibility that *p* exercises the same capacity he exercises when straightforwardly judging that *p*, but with extra sophistication. Let us note here also that it is part of this claim that complexities of conceptual content must manifest themselves in parallel ways in both places. If judging that *p* is likely to lead on to judging that *q* (because the one entails the other and the thinker is interested in whether or not *q*), then wondering 'What if *p*?' will probably lead to the judgement 'If *p* then *q*'. If it did not then what could have justified us in attributing the content 'What if *p*?' to the initial question?[5]

Simulationists have used the case of thinking about possibilities on our own behalf (e.g. when we try to get clear about our various possible courses of action and their outcomes) to illustrate their view of what goes on when we think about others (see, e.g., Heal, 1986; Goldman, 1989). And the 'unhooking' and 'running off line' imagery has been used here too. When I wonder 'What if *p*?' this, it is suggested, leads me to unhook my reasoning machinery and feed in the pretend premise

[5] A simulationist need not be hostile to some broadly functionalist story about the relations between various categories of psychological state, e.g. belief, desire, supposition. Consequently she can recognize and welcome constraints of the kind pointed to here, e.g. that something could not be the belief that *p* unless it was capable of bringing to an end attempts to make it the case that *p*, or that a state cannot count as wondering 'what if *p*?' unless it leads to conditional beliefs which mirror actual inference from the belief that *p*.

that *p*; the machinery then goes through the evolutions it would if I actually believed *p* and were trying to work out its consequences.

But how exactly is this hypothesis meant? Considering this will bring out the difference between the two approaches to simulation which I have been contrasting. It looks as if it is an empirical question whether hypothetical thought is done by actually re-using some of the cognitive structures used in non-hypothetical reasoning. It is certainly imaginable that we discover that different areas of the brain are active in hypothetical and non-hypothetical thinking. And if it turned out that different areas were active then this might constitute some kind of difficulty for the 'off line use' hypothesis — although whether it would be a conclusive objection is extremely obscure since it is a very tricky matter to set out the identity criteria for such things as 'cognitive structures'. But no developments of this kind could constitute evidence that people had two separate theories, one about actuality and one about possibility, because we can make no sense of the idea.

So we come finally to the view of simulationism which I would like to propose. It says the same about thought about other people's thoughts as we have just said of thought about possibilities. We can imagine a thinker who is incapable of thoughts about others' thoughts. But if a person does have the capacity to think about others' thoughts then exercising it will involve exercising the capacity to think about the subject matter of those thoughts, together with some extra sophistication. On this conception of the shape of the simulationist hypothesis further interesting questions will then arise, concerning the nature of this extra setting. Gordon seems to be interested in pressing as far as possible the idea that the extra is very minimal and does not consist of anything like further concepts or a theoretical framework. But another line would be to concede quite a lot to the theory theorist at this point, by allowing that the extra sophistication consists in grasp of general notions like 'belief', 'perception', 'feeling', 'desire', 'action', etc. and of some premises about the kinds of interaction they enter into.[6] To concede this is, however, very far from conceding the whole of the theory theorist's picture, because the way that particular predictions are arrived at, on the basis of information about particular thoughts, is *not* taken to be done by a theory. On the contrary, it is at this point that simulation is necessary. There is a crucial difference between, on the one hand, allowing that people who think about

[6] See note 3 above.

others' thoughts know such generalities as that beliefs and desires tend to lead to action and, on the other, allowing that they have some theory which shows what the specific beliefs that *p, q* and *r* will lead to given the specific desire that *s*. Arriving at such a particular prediction, says the simulationist, can only be done by actually entertaining oneself the thoughts *p, q, r* and *s* and thinking through their implications and interconnections.

To summarize, then, the proposal is that the central differences between a theory theorists and a simulationist (on the second reading of the issue) is that the former maintains that the capacity to think about thoughts is separable from the capacity to think about their objects, while the latter takes it that the capacity to think about thoughts must be seen as an extension of the capacity to think about their objects.

But how does this way of looking at the matter prevent the collapse which happened on the earlier conceptualization? That proceeded by a kind of ju-jitsu move, where the truth of the simulationist story was conceded and then shown to imply the correctness of the theory theorist's claim. Can we repeat the manoeuvre? We cannot because there is no route, never mind what definition we have of tacit knowledge, from the premise that one capacity is an extension of another to the conclusion that they are separate.

Some uneasiness, however, may linger on the grounds that we seemed earlier to have some extremely general considerations which showed that any complex intellectual ability, such as that of predicting others' thoughts, must be such that we can plausibly view it as manifesting grasp of a tacit theory. Surely there will be patterns of dependence and co-variance which will warrant postulation of some causal structure mediating the inferences and that will in turn warrant postulation of a tacit theory? This however is too rapid. The earlier discussion proceeded on the assumption, made only for the sake of the argument, that there did exist a humanly knowable theory of thinking capable of delivering not merely general truisms about connections of beliefs, desires and actions but the sort of specific predictions spoken of two paragraphs ago. Given that assumption, then the argument from competence in psychological prediction to tacit knowledge of the imagined theory goes through. But it may well be part of the simulationist's case, when fully made out, that there is no such humanly knowable specific theory of thinking. And if this can be made good (which is something I shall not attempt here) then the argument lapses.

I have said nothing as yet about why, on this second conceptualization of the debate, we ought to find the simulationist view congenial. Relatedly I have said nothing to show that the theory theory, in its full blooded form, is unattractive. And full exploration of these issues goes beyond the remit of this paper. But I shall offer some brief remarks. The key to both issues is, I would suggest, the fact that thoughts have content, i.e. they represent the world, and that it is in virtue of what they represent that they have their identity as thoughts, and hence their explanatory roles. Content, its immense variety and complexity and hence the difficulties of dealing with it theoretically, have not yet received serious attention in this particular dispute about simulation and theory, where the examples of psychological understanding of others presented by theory theorists are invariably extremely trivial and schematic. All the difficulties that manifest themselves in actual attempts to provide theoretical accounts of inference and decision (e.g. the so-called Frame Problem and the related difficulties of fitting in *ceteris paribus* conditions and explaining the role of background knowledge) have so far not figured at all largely in this debate.[7] This could with advantage be remedied. It is, for example, worth noting that the theory theorist is supposing not merely that there exist graspable, finitely statable solutions to these problems but that we already, tacitly, know what they are, and we know them in some more substantive sense than that we are actually capable of doing the thinking in which we take appropriate account of background conditions, know what to do when the other things are not equal and so forth. This is a remarkable achievement, especially as we are supposed to master the theory which handles these matters at the mother's knee.

A further, and perhaps even more fundamental, line of thought, at which I would like to gesture, is that of externalism about content. The theory theorist, as sketched here, is committed not only to the idea that thoughts are one thing and their objects another (which no one would deny) but also to the idea that we can, in *some* sense, think about thoughts without thinking about their objects. This seems in turn to commit the theory theorist to identifying thoughts as entities (and explaining all their important roles in psychological explanation) by reference to properties of a non-semantic character, for example, quasi-

[7] For a brief but accessible account of the frame problem see Dennett (1984), which is also reprinted in Boden (1990). This latter collection also contains a number of other papers relevant to this question.

syntactic properties or 'senses', where 'senses' can be characterized in a strongly internalist manner. The theory theorist claims not only that we might in some imaginable future neurological or cognitive theory characterize thoughts in this way but that we already tacitly do so. The simulationist, by contrast, will contend that there is at present no way of thinking about thoughts except as semantically characterized and no way of understanding their explanatory role except in terms of the contribution they make to rational intelligibility in virtue of that semantic content. The really thorough going simulationist will (as I have briefly hinted above) urge that there are insuperable difficulties of principle in imagining future scientific theories (in cognitive science, neurology or whatever) which could remove the need to simulate. So, it might be argued, thoughts are of their nature unamenable to fully comprehensive theorizing of the natural scientific kind. Although I am sympathetic to this more ambitious thesis, I do not claim to have said enough here to have made it plausible.[8]

If the issues just mentioned are central to the theory theory vs. simulation dispute then we can see why it is potentially extremely important in philosophy of mind. We can see also how, on this construal of the dispute, it is continuous with earlier discussions (e.g. of Verstehen) and how it is bound up with broader questions about the similarities and differences between the natural and the human sciences.

[8] See Heal (1986 and in press) for some further consideration on these issues.

Bibliography

Allison, H. E. 1983: *Kant's Transcendental Idealism.* New Haven: Yale University Press.

Baron-Cohen, S., Leslie, A. M. and Frith, U. 1985: Does the autistic child have a 'theory of mind'? *Cognition*, 21: 37–46.

Bennett, J. 1966: *Kant's Analytic.* Cambridge: Cambridge University Press.

Bisiach, E., Berti, A. and Vallar, G. 1985: Analogical and logical disorders underlying unilateral neglect of space. In M. Posner and O. Marin (eds), *Attention and Performance*, vol. 11. Hillsdale, New Jersey: Erlbaum.

Bisiach, E., Geminiani, G., Berti, A. and Rusconi, M. L. 1990: Perceptual and premotor factors of unilateral neglect. *Neurology*, 40: 1278–1281.

Bisiach, E. and Luzzatti, C. 1978: Unilateral neglect of representational space. *Cortex* 14: 129–133.

Bisiach, E. and Vallar, G. 1988: Hemineglect in humans. In P. Boller and J. Grafman (eds), *Handbook of Neuropsychology*, vol. 1. Amsterdam: Elsevier.

Boden, M. 1990: *The Philosophy of Artificial Intelligence.* Oxford: Oxford University Press.

Brewer, B. 1992: Unilateral neglect and the objectivity of spatial representation. *Mind and Language*, 7: 222–239.

Cassam, Q. 1987: Transcendental arguments, transcendental synthesis, and transcendental idealism. *Philosophical Quarterly*, 37: 355–378.

Cassam, Q. 1989: Kant and reductionism. *Review of Metaphysics*, 43: 72–106.

Cassam, Q. forthcoming: Transcendental self-consciousness. In P. K. Sen and R. Verma (eds), *The Philosophy of P. F. Strawson.*

Chisholm, R. 1981: *The First Person.* Brighton: Harvester Press.

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

Coltheart, M. 1980: Deep dyslexia: a right-hemisphere hypothesis. In M. Coltheart, K. Patterson and J. C. Marshall (eds), *Deep Dyslexia*, pp. 326–380. London: Routledge and Kegan Paul.

Davidson, D. 1984: What metaphors mean. In *Inquiries into Truth and Interpretation*, 245–264. Oxford: Oxford University Press.

Davies, M. 1986: Tacit knowledge, and the structure of thought and language. In C. Travis (ed.), *Meaning and Interpretation*, 127–158. Oxford: Blackwell.

Davies, M. 1987: Tacit knowledge and semantic theory: Can a five per cent difference matter? *Mind*, 96: 441–462.

Davies, M. 1989: Tacit knowledge and subdoxastic states. In A. George (ed.), *Reflections on Chomsky*, 131–152. Oxford: Blackwell.

Dennett, D. 1984: Cognitive wheels: the frame problem of AI. In C. Hookway (ed.), *Minds, Machines and Evolution*, 129–151. Cambridge: Cambridge University Press.

Dennett, D. C. 1991: *Consciousness Explained.* Boston: Little, Brown.

Diamond, S. 1972: *The Double Brain.* London: Churchill Livingstone.

Evans, G. 1973: The causal theory of names. *Proceedings of the Aristotelian Society*, supp. vol. 47: 187–208.

Evans, G. 1981: Semantic theory and tacit knowledge. In S. Holtzman and C. Leich (eds), *Wittgenstein: To Follow a Rule*, 118–137. London: Routledge and Kegan Paul. (Reprinted 1985 in *Collected Papers*, 322–342. Oxford: Oxford University Press.)

Evans, G. 1982: *The Varieties of Reference*, ed. J. McDowell. Oxford: Oxford University Press.

Fogelin, R. 1985: *Hume's Skepticism in the* Treatise of Human Nature. London: Routledge & Kegan Paul.

Gallistel, C. R. 1980: *The Organization of Action: A New Synthesis.* Hillsdale, New Jersey: Erlbaum.

Gallistel, C. R. 1990: *The Organization of Learning.* Cambridge, Mass.: MIT Press.

Gazzaniga, M. 1988: In A. J. Marcel and E. Bisiach (eds), *Consciousness in Contemporary Science*, 226ff. Oxford: Oxford University Press.

Gibson, J. J. 1979: *The Ecological Approach to Visual Perception.* Boston: Houghton Mifflin.

Goldman, A. I. 1989: Interpretation psychologized. *Mind and Language*, 4: 161–185.

Goldman, A. I. 1992: In defense of the simulation theory. *Mind and Language*, 7: 104–119.

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

Gopnik, A. and Wellman, H. 1992: Why the child's theory of mind really *is* a theory. *Mind and Language*, 7: 145–171.

Gordon, R. M. 1986: Folk psychology as simulation. *Mind and Language*, 1: 158–171.

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

Gordon, R. M. 1992b: Reply to Stich and Nichols. *Mind and Language* 7: 87–97.

Gordon, R. M. 1992c: Reply to Perner and Howes. *Mind and Language*, 7: 98–103.

Gordon, R. M. in press: Simulation without introspection or inference from me to you. In M. Davies and T. Stone (eds), *Mental Simulation: Philosophical and Psychological Essays.* Oxford: Blackwell.

Harris, P. L. 1989: *Children and Emotion: The Development of Psychological Understanding.* Oxford: Blackwell.

Harris, P. L. 1991a: The work of the imagination. In A. Whiten (ed.), *Natural Theories of Mind: The Evolution, Development and Simulation of Everyday Mindreading*, 283–304. Oxford: Blackwell.

Harris, P. L. 1991b: Letter to Josef Perner, 30 May 1991.

Harris, P. L. 1992: From simulation to folk psychology: the case for development. *Mind and Language*, 7: 120–144.

Heal, J. 1986: Replication and functionalism. In J. Butterfield (ed.), *Language, Mind and Logic*, 135–150. Cambridge: Cambridge University Press.

Heal, J. in press: How to think about thinking. In M. Davies and T. Stone (eds), *Mental Simulation: Philosophical and Psychological Essays*. Oxford: Blackwell.

Hurley, S. L. in preparation: *The Reappearing Self*.

Jeeves, M. A. 1965: Agenesis of the corpus callosum — physio-pathological and clinical aspects. *Proceedings of the Australian Association of Neurologists*, 3: 41–48.

Johnson-Laird, P. N. 1983: *Mental Models*. Cambridge: Cambridge University Press.

Kant, I. 1933: *The Critique of Pure Reason*. Tr. Kemp Smith, N. London: Macmillan.

Lockwood, M. 1989: *Mind, Brain and the Quantum: The Compound 'I'*. Oxford: Blackwell.

Marcel, A. J. 1993: Slippage in the unity of consciousness. In Ciba Foundation Symposium No. 174, *Experimental and Theoretical Studies of Consciousness*. Chichester: John Wiley.

Marks, C. E. 1981: *Commissurotomy, Consciousness and the Unity of Mind*. Cambridge, Mass.: MIT Press.

Milner, A. D. and Jeeves, M. A. 1979: A review of behavioural studies of agenesis of the corpus callosum. In I. S. Russell, M. W. Van Hof and G. Berlucchi (eds), *Structure and Function of Cerebral Commissures* 428–483. London: Macmillan.

Nagel, T. 1979: Brain bisection and the unity of consciousness. reprinted in T. Nagel, *Mortal Questions*. Cambridge: Cambridge University Press. (First published in 1971 in *Synthese*, 20.)

O'Keefe, J. 1985: Is consciousness the gateway to the hippocampal cognitive map? A speculative essay on the neural basis of mind. In D. A. Oakley (ed.), *Brain and Mind*, 59–98. London: Methuen.

O'Keefe, J. 1990: A computational theory of the hippocampal cognitive map. In J. Storm-Mathisen, J. Zimmer and O. P. Ottersen (eds), *Progress in Brain Research*, 83: 301–312. Amsterdam: Elsevier.

O'Keefe, J. 1991: The hippocampal cognitive map and navigational strategies. In J. Paillard (ed.), *Brain and Space*, 273–295. Oxford: Oxford University Press.

O'Keefe, J. 1993: Kant and the sea-horse. In N. Eilan, B. Brewer and R. McCarthy (eds), *Spatial Representation: Problems in Philosophy and Psychology*, 43–64. Oxford: Blackwell.

O'Keefe, J. and Nadel, L. 1978: *The Hippocampus as a Cognitive Map*. Oxford: Oxford University Press.

Parfit, D. 1984: *Reasons and Persons*. Oxford: Oxford University Press.

Peacocke, C. 1986: Explanation in computational psychology: language, perception and level 1.5. *Mind and Language*, 1: 101–123.

Peacocke, C. 1989: When is a grammar psychologically real? In A. George (ed.), *Reflections on Chomsky*, 111–130. Oxford: Blackwell.

Peacocke, C. 1992: *A Study of Concepts*. Cambridge, Mass.: MIT Press.

Peacocke, C. 1993: Externalist explanation. *Proceedings of the Aristotelian Society*, 93: 203–230.

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

Perner, J. and Howes, D. 1992: 'He thinks he knows': and more developmental evidence against the simulation (role taking) theory. *Mind and Language*, 7: 72–86.

Piaget, J. and Inhelder, B. 1951/1975: *The Origin of the Idea of Chance in Children.* New York: Norton.

Powell, C. T. 1990: *Kant's Theory of Self-Consciousness.* Oxford: Oxford University Press.

Quine, W. V. O. 1960: *Word and Object.* Cambridge, Mass.: MIT Press.

Rorty, R. 1970: Strawson's objectivity argument. *The Review of Metaphysics*, 24: 207–244.

Schwyzer, H. 1990: *The Unity of Understanding.* Oxford: Oxford University Press.

Sergent, J. 1990: Furtive incursions into bicameral minds. *Brain*, 113: 537–568.

Seymour, S., Reuter-Lorenz, P. and Gazzaniga, M. 1994: The disconnection syndrome: basic findings reaffirmed. Abstracted in The Society of Neuroscience, 1993.

Shebilske, W. L. 1984: Context effects and efferent factors in perception and cognition. In W. Prinz and A. F. Sanders (eds), *Cognition and Motor Processes.* Berlin: Springer-Verlag.

Shoemaker, S. 1984: Causality and properties. In S. Shoemaker, *Identity, Cause and Mind.* Cambridge: Cambridge University Press.

Sperry, R. W. 1990: Forebrain commissurotomy and conscious awareness. In C. Trevarthen (ed.), *Brain Circuits and Functions of the Mind.* Cambridge: Cambridge University Press.

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

Stich, S. and Nichols, S. in press: Second thoughts on simulation. In M. Davies and T. Stone (eds), *Mental Simulation: Philosophical and Psychological Essays.* Oxford: Blackwell.

Strawson, P. F. 1959: *Individuals.* London: Methuen.

Strawson, P. F. 1966: *The Bounds of Sense.* London: Methuen.

Tegnèr, R. and Levander, M. 1991: Through a looking glass. *Brain*, 114: 1943–1951.

Trevarthen, C. 1974: Analysis of cerebral activities that generate and regulate consciousness in commissurotomy patients. In S. Dimond and J. G. Beaumont (eds), *Hemisphere Function in the Human Brain.* London: Elek Science.

Trevarthen, C. 1984: Biodynamic structures. In W. Prinz and A. F. Sanders (eds), *Cognition and Motor Processes.* Berlin: Springer-Verlag.

Walker, R. 1978: *Kant.* London: Routledge.

Wiggins, D. 1980: What would be a substantial theory of truth? In Z. van Straaten (ed.), *Philosophical Subjects: Essays Presented to P. F. Strawson*, 189–221. Oxford: Oxford University Press.

Wilkie, D. M. and Palfrey, R. 1987: A computer simulation model of rats' place navigation in the Morris water maze. *Behavioural Research Methods, Instruments and Computers*, 19: 400–403.

Williams, B. 1978: *Descartes: The Project of Pure Inquiry.* Harmondsworth: Penguin.

Wilson, M. D. 1987: *Descartes.* London: Routledge & Kegan Paul.

Wimmer, H. and Perner, J. 1983: Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13: 103–128.

Wimmer, H., Hogrefe, G.-J. and Perner, J. 1988: Children's understanding of informational access as a source of knowledge. *Child Development*, 59: 386–396.