The Philosophical Quarterly Vol. 44, No. 175 *April 1994*
ISSN 0031-8094

INTERPRETATION THEORY AND THE FIRST PERSON

By Richard Moran

I. INTENTIONAL PSYCHOLOGY AS A THEORY

There has been something close to a consensus in contemporary philosophy of mind that our everyday talk about beliefs and desires, etc., involves the application of a kind of *theory* to what people do. As Ruth Garrett Millikan has recently put it, 'At the present time, a premise shared by many who agree on little else is that ordinary talk about thoughts and other "private episodes" constitutes a sort of folk theory, a primitive science of psychology, the purpose of which is to predict and explain behaviour'.[1] On the analogy with 'folk medicine' and other 'folk wisdom', this theory is sometimes called 'folk psychology', even by some of those who do not endorse the pejorative connotations of the analogy. In this paper I want to look at the 'theory theory' of commonsense psychology as a global account of psychological discourse that gives an account of both the first- and third-person uses of psychological terms in a unified manner. For the most part, the discussion will be restricted to one broad version of this theory, the theory of 'rationalizing interpretation', as it is found in the writings of Davidson and Dennett. There are important differences between their two positions, to be sure; but there is something like a general schema shared by both of them, which may be called Interpretation theory, and this schema has been widely influential, not only among those who adhere to the specific views of either Davidson or Dennett. This paper will first sketch the barest outline of the rationalizing view of psychological discourse, and then say something about some requirements which I think must be satisfied by any adequate account of commonsense psychology, specifically having to do with accommodating both first- and third-person ascriptions of mental states. I then look at one defence of the application of Davidsonian interpretation to the first-person case. Problems with such a picture have suggested to many people that we should move away from rationalizing interpretation as an account of belief-attribution, and move towards an account based on 'cognitive simulation', projecting ourselves into the state of mind of the other person and determining what *we* would think in those circumstances.

[1] *PR* 1986, p. 47; full bibliographic references at end.

It will be argued that Interpretation theory and Simulation theory are not rightly understood as in competition with each other, but are two mutually dependent aspects of the general picture presented in Interpretation theory. There is a *sense* in which Interpretation theory promises to capture both the first- and third-person uses of psychological terms, but this is not because awareness of one's own beliefs generally involves application of an essentially third-person theory to oneself. For Interpretation theory to capture the special features of the first person it must detach itself from the assumption that the defining point of the conceptual network of commonsense psychology is the explanation and prediction of behaviour. References to belief and other intentional states perform distinct but complementary roles in first- and third-person contexts. The assumption of rationality in Interpretation theory makes a tacit commitment to this duality of roles, but the essentially pragmatic understanding of the point of commonsense psychology obscures this and makes the first-person case seem more anomalous than it is.

Interpretation theory is presented as a rational reconstruction of the ordinary practice of psychological attribution. Dennett defends what he calls 'The Intentional Stance', not just as a useful strategy for the explanation and prediction of behaviour, but as a network of principles that, in some sense, provides the meaning of ordinary psychological terms ('Making Sense of Ourselves' p. 84). The principles that define states from within the Intentional stance are the same as (or perhaps idealizations of) the principles that define psychological states within ordinary discourse. Similarly, Davidson has argued that the concepts of belief and desire are defined by their role in the general interpretative strategy of making sense of what people do, where 'making sense' means fitting their behaviour into a pattern of beliefs and desires that render that behaviour rational, given the circumstances. This is to say not only that intentional states such as belief are subject to rational constraints of coherence and consistency, but that the very identity of such states is given by the role they play in such 'rationalizing' interpretations of behaviour. Outside such an explanatory project, we simply have no concepts of belief or desire. On this view, ascribing beliefs to a person is not a purely descriptive affair, but is guided by assumptions about what the person *ought* to believe. Past a point of apparent irrationality in some system of beliefs, we lose the sense and point of speaking of beliefs or other intentional states at all. Related to this is the idea that

beliefs are organized into a *system*, and the rationality of a single belief cannot be evaluated except in its relation to the other beliefs of the system. Hence within the situation of interpretation we cannot ascribe beliefs piecemeal to a person, but must make and adjust our ascriptions in terms of holistic systems of belief.

II. SELF-ASCRIPTION AND UNIVOCALITY

Mental states are identified in quite different ways, depending on whether it is one's own states or those of someone else that one is identifying. In the normal case, a person does not observe his own actions to learn of his beliefs and desires, nor is his awareness of them in the service of an explanation or prediction of his own behaviour. None the less, *self*-ascriptions of mental states are made within the same framework and vocabulary of commonsense psychology as are third-person attributions. It might be asked, then, how it is that a person has non-observational awareness of just those same beliefs which would 'rationalize' his behaviour. Coming from such different directions, how do the interpreters and the agent himself so often end up ascribing the very same beliefs to him? A good account of the conceptual structure of commonsense psychology should satisfy two demands that appear to pull in two different directions. First, it should preserve the assumption that terms like 'belief' and 'desire' have the same meaning when used in first-person and in third-person contexts. This is a problem for the 'theory theory' in general, in so far as the meanings of the terms of 'folk psychology' are taken to be fixed by their role in the theoretical network.[2] Without this assumption of the univocality of psychological terms, we could not say that two people could be talking about the same thing, or even the same *sort* of thing, when, as we say, they talk together about the beliefs of one of them. Secondly, our account should be able to accommodate the *prima facie* differences there are in the manner of first- and third-person psychological ascriptions, for these differences seem just as essential to what we think of as ordinary psychological states as does their involvement in rationality. That is, we want univocality across first- and third-person contexts, but not at the cost of a theory which denies that a person becomes aware of his own beliefs and desires in a

[2] 'If folk psychology is literally a theory ... then the meanings of our psychological terms should indeed be fixed as the thesis of this section says they are: by the set of folk-psychological laws in which they figure. ... The dominant, and perhaps the only, source of meaning for psychological terms is the commonsense theoretical network in which they are embedded' (Churchland, *Matter and Consciousness* pp. 59–60).

manner that is fundamentally different from how he learns of the beliefs and desires of others.

It is possible, of course, to deny either the univocality of psychological terms or the asymmetries of their application in practice, and thus avoid any tension there may be in accommodating both of them. But neither denial is very attractive. The denial of univocality would either be a form of solipsism (e.g., 'when other people speak of "pain" they do not refer to the sort of thing I am feeling now'), or it would be part of an argument that first-person uses of terms such as 'believe' are not part of psychological statements at all. On this latter option, saying 'I believe that my car is parked outside' cannot be a statement about one's own or anyone else's *beliefs*, but is rather just a mode of presenting the contained assertion about one's car.[3] Certainly there *is* such a 'presentational' use of 'believe' and related terms, but sometimes we really mean to speak about our states of mind themselves, and when we do, we use the same terms to describe ourselves psychologically as we use in describing others. The denial of the asymmetries in application has also had its takers, notably in the heyday of logical behaviourism, but also among some philosophers today.[4] The only appeal of this view that I can see depends on confusing it with a denial that the first-person position is special in delivering *infallible* reports on one's state of mind. But one may easily give up the radical epistemic claims for first-person reports, yet still maintain that they are made on a very different basis from reports on the mental life of others. Interpretation theory, or any other global account of commonsense psychology, should be able to account for such asymmetries in application in a way which does not contradict the assumption of the basic univocality of psychological terms.

III. INTERPRETATION THEORY AND THE FIRST PERSON

I take it that the *general* picture of intentional states as being part of a holistic rational system has a good deal of independent plausibility. It does seem essential to anything we could call belief that it must figure

[3] This option is explored by Wittgenstein, but is, I believe, ultimately rejected by him as a general account. It is also discussed in J.L. Austin's paper 'Other Minds'; and is explicitly endorsed in J.O. Urmson's paper 'Parenthetical Verbs'.

[4] Ryle in *The Concept of Mind* is probably the most familiar instance; but more recently, for example, ch. 4 of Paul Churchland's *Matter and Consciousness* is guided by this assumption. See also Richard Rorty for the claim that 'our knowledge of what we are like on the inside is no more "direct" or "intuitive" than our knowledge of what things are like in the "external world" ' (*Synthese* 1982, pp. 330–1).

in explanations of behaviour in something like the way described by Davidson and Dennett, and that there are normative constraints on how unhinged from rationality we can coherently allow a system of beliefs to be. But this is, of course, an essentially third-person point of view on intentional states and their attributions. In an exposition of Davidson's account of mental states, Michael Root (p. 294) stresses the third-person aspect of Interpretation theory, and the sense in which its normativity appears to give the theory a constituting role in determining the nature of types of intentional states, as well as in individual attributions:

> For both Quine and Davidson, we owe our idea of the mental to our interest in explaining the behaviour of others, and, for both, our idea of the mental is constituted by the way that we pursue that interest: we offer a rational explanation of the behaviour.

Root sees that if this is, as it surely appears to be, a claim about the very concepts of intentional psychological states, then it must provide a way to describe their first-person as well as third-person applications.[5] If even the general outline of Interpretation theory is right, then it may seem either that there must be two distinct concepts of the mental, which seems unacceptable, or that there must be a way to see the *self*-attribution of intentional states as part of the same interpretative project of the third-person perspective. If one claims, however, that psychological terms get whatever meaning they have from their role in the explanatory theory of intentional psychology, then there will be considerable pressure somehow to assimilate the case of self-attribution to the paradigm of understanding others.

On one understanding of it, Interpretation theory appears to promise a way of explaining both why the two routes to the ascription of mental states normally agree in their results (i.e., in their particular attributions), and how the meaning of the terms for intentional states can remain uniform across first- and third-person contexts. A possible explanation would proceed from the claim that there is but one concept of an intentional mental state, and it is the one employed in third-person rationalizing interpretations of behaviour. Despite appearances to the contrary, this same process of interpretation is applied to oneself in the case of self-attribution of beliefs or other mental states, and thus uniformity of meaning is not threatened. The two types of attribution agree in their results because they are at bottom the same: 'introspection' is not direct

[5] Naturally, both Quine and Davidson would eschew all such talk of *conceptual truths*, but that does not alter the fact of the global and *a priori* nature of their claims about psychological discourse.

acquaintance, but is just another form of rationalizing interpretation. Within the Davidsonian perspective, Root considers the possibility that the self-attribution of psychological states follows the same charitable and holistic principles as does the rationalizing interpretation of the behaviour of others. The following passage (p. 298) gives a sense of how such essentially third-person principles of interpretation might be followed in the first-person case:

> When I attribute a belief to myself, the aptness of my attribution is judged by the same norms that I use to judge the aptness of my attributions of belief to others.... What grasp I have of the content of one of my beliefs, I have in virtue of what I take to be the content of other beliefs of mine. If it comes to my attention that a body of my beliefs is not coherent, I have to give something up. I could give up the beliefs, but I have another alternative: I can give up my understanding of them. That is, the fact that my beliefs, as I understand them, are inconsistent, is a reason to change the beliefs, but it is equally a reason to change the way I understand them. In short, charity must guide me when I try to understand others, but it must guide me when I try to understand myself as well. Charity does not begin with other minds: it begins at home.

On this view, when I discover an apparent contradiction in my own beliefs, I have the same interpretative options before me as in the case of interpreting the behaviour of someone else. If the beliefs I attribute to the other seem radically false, I may 'charitably' preserve rationality by revising my original *attribution* of belief to bring it in line with the facts as I see them. At this point I do not mean to quarrel with the details of this story, for I am interested in this as an important case for any account of commonsense psychology that takes it to involve the application of a theory. It is likely that any version of the 'theory theory' will make some appeal to rationalizing explanation, and any version of that theory will have to account for *self*-ascriptions in some way.

In assessing the plausibility of this picture, we must remind ourselves of the dual aspect it claims for ordinary psychological ascriptions. Within Interpretation theory, decision theory is usually taken to offer one paradigm formalization of the rational interaction of beliefs and desires, a formalization that captures the structure of our ordinary understanding of them as psychological states. And decision theory itself can be looked at both descriptively and prescriptively, both as a description of what people do in choice situations, and as a rational prescription for what they ought to do (see Root p. 279). Now the application of this dualism to the case of self-interpretation can be seen by looking at a distinction between two different sorts of questions a person may ask about his current attitudes

and mental life. When a person asks himself the question, 'What do I think about X?', this may be a purely theoretical question enquiring into what his present belief about something is; or it may be the question he asks himself in the course of making up his mind about something, in which case he is asking what we may call a deliberative question about what he *is to* believe about that thing. The theoretical question is answered by a discovery (or description) of what he antecedently does believe, while the deliberative question is answered by a decision (or prescription) about what to believe.[6] When Root speaks in the passage quoted of applying Interpretation theory to oneself, the wording suggests a situation in which the person is considering what antecedent belief or desire to attribute to himself. This would be to view the question he is asking as a purely theoretical question about his belief or desire. But genuine self-reflection normally involves interaction between the two kinds of question in such a way that it is indeterminate when a putatively theoretical question is in fact being answered as a deliberative question about what *to* believe. In the situation of self-interpretation, the descriptive and the prescriptive are entwined with each other in much the same way as they are in decision theory's account of the interaction of belief and desire.

If this shifting between the normative and the descriptive in the language of intentional psychology is something that obtains consistently across both first- and third-person contexts, then it may be possible to detach the strategy of interpretation from the Instrumentalist slant which it is usually given by its defenders. Consider Dennett's well-known example (in 'Intentional Systems') of the attempt to predict the behaviour of a chess-playing computer, as a case of the essentially third-person perspective of Interpretation theory. It may well be possible in principle to infer the computer's future chess-playing behaviour from either the Physical or Design stances, but even if we could master the necessary concepts they would be clearly too cumbersome to apply in actual practice. Within the Intentional stance we are already familiar with the requisite concepts, for they are the familiar ones of ordinary psychology (with the addition of a knowledge of chess). And there is also the following significant feature of the Intentional stance: only from within *this* perspective do I determine what the computer will do in a particular situation by determining what I would (or should) do in that situation. Having made the guiding assumption of rationality dictated by the Intentional stance, my only remaining question is: what *would* I do in that situation? And this is now a deliberative question about how to solve a particular chess problem, and not a theoretical question about

[6] This distinction is developed in my 'Making up your Mind', *Ratio* 1988.

future behaviour. From within this stance, the only way I can answer the predictive question about what the computer will do is by answering the non-predictive, normative question about what I would do.

The deliberative question of what I *am* to believe about X is normally answered on the basis of consideration of the facts about X, not about myself. We can describe this by saying that the self-addressed deliberative question 'Shall I believe that p?' is *transparent* to the theoretical question 'Is it the case that p?', which is not about oneself. 'Transparency' here means that the only way I have for answering the former question (what *to* believe) is by attending to whatever would answer the latter (what is the case). And it will ordinarily (though not always) be the case that the theoretical question about one's current belief is transparent to a theoretical question about the world, the object of one's belief. This situation is duplicated in Dennett's case of predicting the chess-playing computer's move, for in searching for an answer there, one's attention will not be focused on the computer, but rather on the *chess problem*. One need not look at (or inside) the computer at all. This shows the third-person, predictive methodology of the Intentional stance to be part of the first-person process of answering the deliberative question of what one should believe or do in such-and-such a situation. From within this stance, any reason I may have for believing the computer will do X is at the same time a reason for believing that I would or should do X in that situation.

This re-working of the normative aspect of Interpretation theory helps explain why first- and third-person attributions normally agree. The principle of charity which guides interpretation is sometimes described in terms of maximizing the rationality of the speaker being interpreted, but it is just as often described in terms of formulating interpretations so as to maximize agreement of the speaker's beliefs with those of the interpreter. On the face of it, these may look like different requirements. As radical interpreters, when we make the beliefs of other speakers come out as close as possible to our own, are we not just assuming that they must have the same false beliefs and unworthy desires as we do? And does that not conflict with the requirement of maximizing the rationality of the speakers? The answer presumably is that our grasp of our own rationality reflects the best grasp of rationality in general that we have. We simply have no choice, in our formulation of charitable interpretations, but to rely on notions of the true, the good, and the reasonable as seen by us. As Davidson has often put it, if we are to interpret behaviour at all, we must appeal to *some* standards of rationality, and which ones can or should we appeal to, if not our own?

Some recent arguments treat this more 'egocentric' interpretation of the principle of charity as a genuine rival to the rationalizing interpretation, and one that is superior to it as a reconstruction of our actual procedures of belief-attribution within intentional psychology. Robert Gordon, Alvin Goldman and others have advanced 'cognitive simulation' as a model for how psychological ascriptions are actually made: what we do is simulate in imagination what we take to be the other person's state of mind, by rehearsing that person's internal and external situation and engaging in deliberation from that altered point of view.[7] One chief advantage claimed for this theory is that it is more realistic as a picture of how people actually do ascribe mental states to others, and it avoids the assumption of idealized rationality that is so prominent a part of Dennett's account of the Intentional stance. What is found hopeful in Simulation theory is that it preserves the general assumption that the concepts of intentional states, and their particular attributions, are defined by their role in explaining and predicting behaviour; but it avoids both the normativity and the implicit instrumentalism of Interpretation theory. There are certainly problems with Dennett's own strategy, and the very broad and unspecific use he makes of the notion of rationality; but I think construing Simulation theory as an independent alternative to the essentials of the rationalizing picture rests on a misunderstanding of the kind of imagining involved in cognitive simulation, specifically on a misunderstanding of the role of the first person in this kind of imagination. Simulation theory does not offer an independent alternative to the rationalizing procedures of Interpretation theory, though it may offer a way to mitigate that theory's exclusively third-person perspective.

Stich is quite explicit about taking Simulation theory not only to avoid Dennett's idealized or uncritical appeals to the concept of rationality, but also to be an entirely non-normative method of belief-attribution. Comparing his account to Dennett's, he says ('Dennett on Intentional Systems', p. 59):

> The second reason for preferring my line to Dennett's ... is that the idea of a *normative* theory of beliefs and desires, which is central to Dennett's view, plays no role in mine. And this notion, I would urge, is one we are best rid of.

[7] See Gordon 1986; Goldman 1989; Heal; Morton; Stich 1981 and 1982. Since these original papers were published, the literature on Mental Simulation has proliferated and the differences between different proponents of the theory have been sharpened. My characterization of the view is meant to be broad enough to capture distinctive features of the core of the idea, but individual authors may feel that some of my claims do not apply to them.

It is not clear that normativity plays no role here. When we put ourselves in someone else's shoes and project the beliefs and desires we would then have onto that person, what reason have we for thinking that this exercise could possibly tell us anything reliable about this other person's beliefs and behaviour? Here is one situation I could imagine: someone drinks a quart of vodka and then tries to drive home at night in the snow. I could project myself into this situation and imagine what I would do, and on this basis I could predict, with some confidence, whether this person will make it home or not. And so far the most salient parts of the story and my interpretation of it are not normatively based. The basis for my confidence in my prediction is, in part, my assumption that this person and I share the same basic physiology. So, Stich might argue, confidence in the psychological ascriptions we make is based on the non-normative assumption that we share the same basic psychological structure.

But we ordinary folk know next to nothing, of a non-normative sort, about our own psychological structure, let alone about the cognitive structures of other people, or of other beings we might describe in the language of commonsense psychology. At the least, we know no more about this than we do about our basic physiology, and yet the psychological projections we routinely make about other people are both more sweeping in scope and more fine-grained in application than even the best physiological knowledge. If Simulation theory is taken alone, removed from any normative rationality assumptions, then we lose the only basis we have for taking cognitive simulation to have any value as a way of understanding others. Within such a 'purified' simulation, one would answer the question about what one would do in some situation simply as a prediction of future behaviour, as in the case of the drunk driver, rather than as a hypothetical decision in favour of it. But normally a person has very little of an inductive, norm-free basis for predicting his own future or hypothetical behaviour. In the ordinary case, involving something more than sheer conjecture, he can answer such first-person questions with confidence only because he can make a here-and-now decision about what course of action to endorse. And without the normative basis he loses the rationale for applying this same prediction to another person, for he cannot assume that whatever inductive basis he had for his first-person prediction would apply to anyone else.

Further, the 'purified' interpretation of projective imagination just seems false to the facts of how we actually treat such productions and what our interest in them is. To compare a related case, when philosophers and others ask themselves about their responses to various

hypothetical situations, they are not making armchair predictions of their future thinking or verbal behaviour. Rather they are presently committing themselves to a particular attitude or decision about the facts. Similarly, the first-person simulation of another person's mental state is typically applied to that other person on a partly normative basis, and not purely as a prediction. Even the imagination of, e.g., someone's irrational fear of the dark will involve the relativized application of normative assumptions concerning possible actions and responses appropriate to such a fear. Without the sense provided by such assumptions, there would be nothing for me to imagine. In answering an 'egocentric' question about what *I* would do or believe in some situation, I avail myself of the same considerations as I would in answering the 'rationalizing' version of the question, which makes no reference to me as a particular person. The kind of interest we do have in the 'egocentric' question is expressed in the fact that we answer it in the same manner as we would answer a deliberative question about what *to* do or believe, rather than as a request for a theoretical prediction about a particular person's future behaviour.

This gives us the beginnings of an explanation for why the rationalizing procedures of Interpretation theory and the egocentric or 'projective' procedures of maximizing agreement will normally arrive at the same results. It also apparently commits the argument to the view that the imaginative simulation of another person's Intentional states is a different sort of affair from the imagination of the sensations of another person. For the account given here of cognitive simulation makes it essentially a matter of reflection, not on internal states, but on the intentional objects of intentional states; and sensations, on most accounts, are not representational states. The two types of imagination *are* essentially different, I believe, but just a few remarks in support of this view will have to suffice here.

A person can pretend that he is in pain, and he can imagine what it would be like to suffer a particular injury and vividly imagine that pain. In cognitive simulation or reflection on thought-experiments, on the other hand, a person does not *pretend* to have a particular belief: rather, he makes up his mind about some hypothetical situation. And that is to form a genuine belief, make a genuine cognitive commitment. Now while it is true that imagining being in pain might occasionally produce real pain for the imaginer, it is surely not intrinsic to such imagining that it must do so. The relevant difference is that a hypothetical pain is just a hypothetical pain (like an imaginary friend), whereas a hypothetical belief is in most situations best seen as a conditional belief, which is a

real belief, just as a conditional intention is a real intention, and involves a genuine cognitive commitment to the world's being a certain way. (For these reasons, I disagree with Goldman[8] when he says that 'a virtue of the simulation theory is its capacity to provide a uniform account of all mental state attributions, not only of propositional attitudes but of non-propositional mental states like pains and tickles'.)

IV. SAVING THE ASYMMETRIES

We may see Simulation theory, then, as a version of, rather than as an alternative to, the basic picture presented in Interpretation theory. But it is none the less a version that points up certain important parallels in the processes of self-ascription and the attribution of intentional states to others. One of the attractions sometimes claimed for Simulation theory is precisely that it gives us a unified picture of attributions within commonsense psychology, since the manner of attribution is the same whether it is applied to oneself or to someone else.[9] However, the fact that the 'egocentric' and the 'rationalizing' versions of the principle of charity coincide in their results stems from facts about a person's relation to his own beliefs which, in other respects, seem to find no parallel in his relation to the beliefs of others. Projecting one's own likely beliefs onto another person is normally equivalent to interpreting him as rational, because it is a logical truth that a person takes his own beliefs to be true. But this same truism makes for very different possibilities in the application of charitable interpretation to others and to oneself. I can take the beliefs I ascribe to another person to be false, and I can see some of my own *past* beliefs as false, but I cannot take any belief I currently hold to be false. The requirement that a person must take his current beliefs to be true is quite different from and stronger than any version of the principle of charity. Even when understood in a strong sense as some requirement of truth (and not just reasonableness) in the beliefs of others, the principle of charity applies without any difference to a person's past and present configuration of beliefs. From the observer's perspective, whatever requirements of rationalizability and interpretability there may be will apply equally to either belief system. By contrast, the requirement that a person view his own beliefs as true does not apply indifferently to his past and present beliefs, as we have seen; hence nothing like the principle of charity accounts for it.

[8] 'Interpretation Psychologized' p. 172.
[9] See Heal p. 149.

A closer look at these differences will show a very different relation to intentional states and their attribution from that suggested by the 'explanation and prediction' model in Interpretation theory. If the egocentric version of the principle of charity (maximizing agreement with the interpreter) functions in essentially the same way as does the rationalizing version, that will mean that the exercise of *self*-rationalizing described by Root will be a matter of securing agreement between what one believes about the world and the self-attributions of belief one is inclined to make: in other words, making one's first- and second-order beliefs cohere. But this is a *very* special kind of coherence, different in aim and procedure from the coherence sought by the radical interpreter. When the outside interpreter discovers that the beliefs he has attributed to some speaker are either internally inconsistent or incompatible with the facts as he knows them, he has several distinct options. (1) He may revise his original attributions of belief to the speaker until the beliefs attributed accord with each other and with the facts. Thus he makes them come out true. But no Interpretation theorist thinks such 'true-making' re-interpretation will always be possible or reasonable within one's overall theory; and so sometimes he will have to take option (2): settle with ascribing beliefs to some speaker which he knows to be false. And finally of course, it is also in principle possible for the interpreter to take option (3), to resolve the conflict by changing his *own* view of the facts, though this possibility is rarely alluded to in the literature on Interpretation theory. This third option would involve the interpreter's keeping the original belief attributions to the speaker fixed and bringing his own beliefs into line with them.

In the situation of self-interpretation, by contrast, *any* type of conflict in attribution calling for resolution will in fact be a conflict between one's first- and second-order beliefs. And in the resolution of *this* type of conflict, the stance represented in (2) is not an option. A person cannot just settle, even disappointedly, with the ascription of false beliefs to himself and leave it at that. To make such a second-order judgement is *ipso facto* to abandon the first-order belief, and with it the original attribution of that belief to oneself. As for option (1), the person cannot simply alter his original *attribution* of false or inconsistent beliefs to himself, where this is understood, as it is in the original description, as something distinct from changing his view of the facts themselves (option 3). Even if Interpretation theory makes the strong claim that beliefs are *constituted* by interpretation, it must preserve the general distinction between, on the one hand, second-order beliefs about, say, what Othello thinks about Desdemona and, on the other, the first-order beliefs which

are about Desdemona, and not about Othello's or anyone else's attitudes. For the outside interpreter, there is nothing clearer than the difference between revising his interpretation of Othello's beliefs (option 1) and changing his mind about Desdemona herself (option 3). And any theory of belief would have to make this distinction somehow. But for Othello himself there cannot be this difference between revising his ideas about what *beliefs* he has about Desdemona and revising his ideas about *Desdemona*. That is, he does not have the option of giving up his original attribution of belief to himself, his second-order belief that he has the belief that Desdemona loves Cassio, without thereby giving up his belief about Desdemona.

The interpreter's simultaneous belief that Desdemona does not love Cassio (first-order) and that Othello believes that she does (second-order) is not treated by him as a *contradiction*. Even within the requirements of charity he realizes that he does not violate any canon of rationality by retaining both beliefs (one about Othello, one about Desdemona). And to be sure, Othello too must realize that the two beliefs together do represent a perfectly possible state of affairs, so that he does not violate any rule of logic in supposing them to be both true at the same time. (The concept of belief is not solipsistic, and anyone credited with the concept of belief is thereby credited with the understanding that the belief that p and the fact represented by p are independent of each other, and that one may obtain in the absence of the other.) But even so, Othello's maintenance of both beliefs cannot be stable; one or the other of them must be given up. They are not in logical contradiction with each other, yet they do systematically conflict with each other, though their counterparts in the mind of the interpreter do not. The believer himself cannot treat his first-order belief as giving him a datum about the world, and the second-order belief as just giving him a datum about the psychology of a particular believer. Rather he must treat questions about *either* of them as transparent to questions about the world. He does not have the option of treating the relation between his first- and second-order beliefs as a conflict between two different belief systems, but only as a conflict within *one* view of the world. By contrast, the rationalizing interpreter does not operate under any such constraint, and this makes the whole notion of self-attribution and the revision of self-attributions a very different matter from the interpretation of others. For entirely different sets of reasons will apply, depending on whether the interpreter is revising his original belief-*ascriptions* (as in option 1), or changing his mind about the *object* of the beliefs in question (option 3). In the first-person position, however, it

is reasons of the latter type which do all the work. Ultimately, any reason the person has for revising his attribution of belief must also function as a reason for changing his mind about the object of belief.

V. RATIONALITY AND THE TWO PERSPECTIVES OF INTENTIONAL PSYCHOLOGY

What do these asymmetries mean for the problem of the univocality of the terms of intentional psychology, as these terms are defined within Interpretation theory? The principles under which first-person belief attributions are made and revised are different from those guiding the ordinary third-person application of the theory, but the two are not unrelated. Third-person ascriptions depend on assumptions concerning the rationality of that person. It follows from this that *explanation* from within the third-person Intentional stance must correlate with *justification* from the first-person stance, otherwise the notion of rationality has no grip on the actual bases of the agent's thought and behaviour. The norms of rationality appealed to in Intentional stance explanations are not like laws of physics to which the object in question is merely *subject*. Rather they are norms to which the subject in question intends his behaviour and reasoning to conform, and Intentional stance explanations themselves require this. Invoking rationality in the first place thus commits the interpreter to a notion of justification, which means justification *from the standpoint* of *the agent and his thought*, rather than from the standpoint of the explanation or the explainers. The matter to which justification considerations must apply is not, in the first instance, the *interpretation* or explanation that is given, but the action or reasoning of the agent himself. And, most importantly, the agent's own justification of his actions must be tied to his capacity to justify the beliefs and desires he has in the first place. He cannot take the interpreter's stance towards himself and simply settle with the attribution of a certain stock of beliefs and desires that best rationalizes his behaviour, whether or not they reflect his best view of how things objectively are. He would not be at all rational if he proceeded this way; and thus, as the interpreter would or should recognize, he would not be a suitable candidate for rationalizing explanation in the first place.

The reasons which *explain* an action are states of mind of the agent, which may themselves be either veridical or mistaken. When a belief that is an explanatory reason is a false belief, this need not affect its explanatory validity in the slightest. But naturally this is not the agent's

own relation to his reasons, which must be *guiding* or justifying reasons, and which are facts distinct from and independent of his beliefs.[10] What provides a reason for betting on a certain horse is the fact that it will win, or evidence for that fact, not one's *belief*, however strong, that it will win. And the *fact* would constitute a reason for the bet, even in the absence of the corresponding belief. A person's own attitude towards how his beliefs and desires relate to his action is not 'subjectivist', nor simply a matter of good fit among the states themselves. His belief is not *for him* a psychological datum which could, even in principle, justify his behaviour purely in its role as a psychological state. Otherwise he would take himself to have as much reason to make the bet whether his belief is true or false. On the other hand, the rationalizing interpreter *will* take that behaviour to be rationalized by the belief, whether it is true or false. The interpreter can afford to treat the belief as a psychological datum, and go on from there to use it in a rationalizing explanation. The agent himself does not have this option, and, as we have seen, he would not be rational if he did. So the interpreter's stance and its success presuppose the stance of the reasoning agent; and further, they presuppose that belief is treated differently from within the two stances.

Does this mean that we are really dealing with two essentially different concepts of belief, so that the threat of ambiguity between first- and third-person psychological discourse now shows up between these two different stances? I think not. In the first place, it is not such a clear matter how one defends a claim as to the unity of a particular concept, in particular how one gives principled answers to questions about how much and what kind of difference in application a unitary concept may tolerate. In this case we are interested in accommodating the assumption that the meanings of psychological terms do not alter across first- and third-person contexts, while at the same time we want to insist on certain fundamental asymmetries in these two contexts. We do not want to show that concepts such as belief do not play more than one role, for it is hard to see how anything remotely like our concept of belief could fail to play a dual role: as explanatory of behaviour and as bearer of truth values. As long as there are believers, and as long as beliefs purport to represent the world, it will be possible to ask of any belief both whether it is true or false, and how it disposes the believer to act. The two aspects of belief depend on each other, since the explanatory aspect requires an assumption of rationality, and *that* requires that the believer himself not treat his belief as an opaque psychological fact. In his awareness of his

[10] See Raz (ed.), *Practical Reasoning*.

beliefs, he either acts upon them as he acts upon recognition of some aspect of the independent world, or he subjects them to revision.

Instead of positing two different concepts of belief, we may speak of different interests in the same concept. From the agent's perspective, the question of the truth of his beliefs is prior to the question of how they will dispose him to act. Beliefs 'aim at truth', and do not enter into his practical reasoning in a way that brackets the question of their truth. The interpreter, on the other hand, will be interested in how beliefs explain behaviour, and this is a role played by false beliefs nearly as often as by true ones. Any representational state will have such a dual aspect, one under which it is transparent to the world in a certain way, another under which it makes a contribution to the behaviour of the agent. Naturally these different interests in belief are not *restricted* to the first- and third-person uses respectively. In communicating and reasoning with others, for instance, we are concerned with the truth, and not just the explanatory adequacy, of the beliefs we take them to have; while, on the other hand, in understanding oneself one will sometimes need to bracket the question of the truth of one's beliefs and concentrate on their explanatory role. And the fact that in the *past*-tense case the person's own relation to his belief may approximate to that of an outside interpreter provides additional reason against thinking of the two roles as involving different concepts. For it is not credible that any shift in meaning occurs when one says (or thinks) something of the form 'I believed it then and I believe it now'.

The general picture of the propositional attitudes in Interpretation theory can *allow* for this dual aspect of belief, and I have claimed that it is in fact tacitly committed to it. But the account it gives of the behavioural basis and explanatory point of psychological ascriptions does not fit the first-person case. And because the first-person stance (that of the reasoning agent) is not eliminable from the core of the rationalizing project described in Interpretation theory, but belongs centrally to it, we cannot take the purely third-person project of explaining and predicting the behaviour of others to define and exhaust the meaning of psychological terms. Explanation and prediction of behaviour is patently not the basis of a person's relation to his own mental life. It is only a somewhat less partial picture of the point of making psychological attributions to others and accepting them from others. This is said on the assumption that Interpretation theory does capture an important part of the conceptual network of commonsense psychology, and can be detached from the implicit behaviourism of Davidson and the explicit instrumentalism of Dennett. It has not been shown here just how this

may be done, but we have seen some reasons to suggest that it can be. These have also included reasons for thinking that we would thereby fit several of the special features of the first-person position into a broader account of the structure of ordinary psychological discourse.

Finally, I hope to have cast some doubt on the 'theory theory' itself, understood as the assumption that the semantic and rational structure of commonsense psychology is that of an explanatory and predictive theory of human behaviour. It is difficult to know how to confront such a general assumption head on, and the few comments here cannot hope to dissuade anyone deeply committed to it. It is also difficult to give a positive description of what the systematic discourse of commonsense psychology would *be*, if not a theory of behaviour. I suspect that the very general agreement with the 'theory theory' among contemporary philosophers is due in large part to the tacit belief that to reject it would be to embrace some form of logical behaviourism, the 'theory theory' having developed as a naturalistic response to the demise of logical behaviourism in the 1960s. But it is not obvious that these are our only alternatives. It seems possible to claim, against logical behaviourism, that there really are persons, agents who act out of beliefs and desires which may outrun the behavioural evidence for them, and at the same time claim that the use of the concepts of belief and desire in one's own reasoning and in many aspects of reasoning and communicating with others is not a theoretical move in the service of explanation and prediction of behaviour. This still leaves it a contingent empirical fact that there are persons at all, and leaves any individual claim about someone's beliefs with the same contingent status.

It is certainly a bad argument against the 'theory theory' that our daily interaction with our acquaintances does not *appear* to us to be a theoretical enterprise. The 'theory theory' is a claim pitched at an extremely high level of abstraction, and at this level many activities within the physical world that do not appear to involve anything like conscious inference can legitimately be seen as having a kind of theoretical basis. (Here we might think of the sort of tacit understanding of the principles of 'folk physics' which tell the child that when he pushes the swing, it will swing back.) The 'theory theory' is not intended to capture the phenomenology of our relations to ourselves and to other people. But, leaving that aside, it may still seem a bad reconstruction of the place and the rationale of the concepts of intentional psychology in our thought and action. We have no more reason to think that we employ the 'theory of persons' in order to explain the motions and noises of the living human bodies around us, than we do to think that we employ the 'theory of persisting physical

objects' in order to explain and make sense of our sensory experience. In both cases the thing that is supposedly inferred, as a result of theoretical explanation, is in fact something given prior to the thing that is presented as our original datum. Physical objects and persons are more primary objects of our desire to understand (including the impulse to theorize) than are either sense-data or non-sentient living human bodies. And the concepts of 'pure experience' and 'pure behaviour' are more theoretically sophisticated and derivative than are the concepts of 'object' or 'person'. To say this is not to beg the question against scepticism with respect to external objects or other minds. If such scepticism were correct, then we who talk about physical objects and persons (and especially those of us who talk *to* persons) would be wrong. But this possibility does not mean that the structure and point of our talk about physical objects or persons has been to explain the behaviour of experiences or non-persons. The fact that some question is an empirical matter does not mean that the whole vocabulary in which it is couched is a theoretical vocabulary.

What is denied, in rejecting the 'theory theory', is the idea that the prediction and explanation of behaviour is the primary and *meaning-constituting* use of the concepts of commonsense psychology. And it follows from this that the survival or replacement of this as a vocabulary for describing persons does not rest with how well it does, compared to some possible rival, as a predictive device. A future theory of behaviour could do very well indeed without providing a reason to eliminate reference to persons and beliefs in our relations to ourselves and to others. For the abandonment of the whole discourse of beliefs, desires and reasons, as applied to human beings, would be the abandonment of the very idea of persons. And that idea, and our engagement with it, is at least as deep and entrenched as the idea that we have any theoretical interests in the first place. Although it is a contingent fact that there are persons, and even if it is strictly conceivable that the whole concept of persons might some day wither away, that eventuality would be an historical development occurring for reasons quite other than theoretical ones.[11]

Princeton University

[11] In writing this paper I have profited from discussions with Sydney Shoemaker, Anthony Appiah and Martin Davies.

REFERENCES

Austin, J. L. 1979: 'Other Minds', in *Philosophical Papers*, 3rd edn (Oxford UP), pp. 76–116.
Churchland, P. 1984: *Matter and Consciousness* (Cambridge, Mass.: Bradford Books).

Davidson, D. 1980: 'Mental Events', and 'Psychology as Philosophy', in his *Essays on Action and Events* (Oxford: Clarendon Press), pp. 207–25, 229–44.
Dennett, D. 1981: 'Intentional Systems', in his *Brainstorms* (Cambridge, Mass.: Bradford Books), pp. 3–22.
—— 1987: 'True Believers', and 'Making Sense of Ourselves', in his The *Intentional Stance* (Cambridge, Mass.: Bradford Books), pp. 13–42, 83–116.
Goldman, A. 1989: 'Interpretation Psychologized', *Mind and Language* 4, pp. 161–85.
—— 1992: 'In Defense of the Simulation Theory', *Mind and Language* 7, pp. 104–19.
Gordon, R. M. 1986: 'Folk Psychology as Simulation', *Mind and Language* 1, pp. 158–71.
—— 1992: 'The Simulation Theory: Objections and Misconceptions', *Mind and Language* 7, pp. 11–34.
Heal, J. 1986: 'Replication and Functionalism', in J. Butterfield (ed.), *Language, Mind and Logic* (Cambridge UP), pp. 135–50.
Millikan, R. G. 1986: 'Thoughts Without Laws; Cognitive Science With Content', *Philosophical Review* 95, pp. 47–80.
Moran, R. A. 1988: 'Making up your Mind: Self-Interpretation and Self-Constitution', *Ratio* NS 1, pp. 137–51.
Morton, A. 1980: *Frames of Mind* (Oxford UP).
Raz, J. (ed.) 1978: *Practical Reasoning* (Oxford UP).
Root, M. 1986: 'Davidson and Social Science', in E. LePore (ed.), *Truth and Interpretation: Essays in the Philosophy of Donald Davidson* (Oxford: Basil Blackwell), pp. 165–89.
Rorty, R. 1982: 'Contemporary Philosophy of Mind', *Synthese* 53, pp. 323–48.
Ryle, G. 1949: *The Concept of Mind* (London: Hutchinson).
Stich, S. 1981: 'Dennett on Intentional Systems', *Philosophical Topics* 12, pp. 38–62.
—— 1982: 'On the Ascription of Content', in A. Woodfield (ed.), *Thought and Object* (Oxford: Clarendon Press), pp. 143–71.
Urmson, J. O. 1952: 'Parenthetical Verbs', *Mind* 61, pp. 480–96.
Wittgenstein, L. 1956: *Philosophical Investigations* (Oxford: Basil Blackwell).