Philosophical Foundations of the Theory Theory of Folk Psychology

Heidi Lene Maibom
University College London

Thesis submitted for the degree of PhD
in philosophy April 2000

Abstract of Thesis

The Theory Theory is one account of our practice of attributing psychological states to ourselves and others. In the late 1980s, the advent of an alternative account, the Simulation Theory, provoked a new debate about the correctness of Theory Theory. In this debate, much confusion has arisen about the nature of Theory Theory. In this thesis, I try to dispel some of this confusion by outlining what must be the philosophical foundations of the Theory Theory.

The thesis makes five main claims. First, I argue that it is unhelpful to regard any body of knowledge whatsoever as a theory. To do so does not illuminate the idea of a theory, and it makes the distinction between Theory Theory and Simulation Theory quite obscure. Second, I argue that we should understand the relevant sense of 'theory' in terms of the idea of a scientific theory. A scientific theory is composed of lawlike generalisations that contain interrelated terms that refer to entities and properties that are explanatory of the data, for example by being causally efficacious in the production of them, or by being related to them in some lawlike fashion. Third, I argue that Theory Theory is not committed to common sense functionalism, in either its semantic or metaphysical versions. Fourth, Theory Theory is not committed to us attributing psychological states to others and to ourselves on the same basis. Hence, it can allow that we can have distinctive knowledge of our own minds. Finally, I argue that Theory Theory should not maintain that we have tacit knowledge of folk psychological theory. From what we know of how this theory looks like, and from what current theories of tacit knowledge hold to be defining features of tacit knowledge, we must conclude that this knowledge is not tacit.

The above conclusions form a basis of the Theory Theory which will make possible more specific formulations of it, and will facilitate and clarify further debate.

Contents

Preface

Like many in my position, I approached my doctoral thesis with high ambitions. As I remember it, I aimed to argue that Theory Theory was the one true account of folk psychology, that it was tacitly known, and that it was innate. As it turns out, I ended up doing none of this. The most important reason was that reading through the Theory Theory versus Simulation Theory debate, it struck me that people were frequently speaking at cross purposes. Simulationists criticised theory theorists for holding this or that view, and vice versa. As far as I could tell, it was sometimes quite unclear that the opponent was wedded to the criticised view. Even worse, sometimes theory theorists disagreed among themselves about what the Theory Theory was committed to, and ditto for the Simulation Theory. So I decided to concentrate on just one of the competing theories.

This is a thesis about the philosophical foundations of the Theory Theory. However, all the issues that I am going to discuss, have been fuelled by the debate between theory theorists and simulationists. A satisfactory formulation of the Theory Theory relies on it being defined in a relatively precise way such that there remains a satisfactory distinction between it and the Simulation Theory. Otherwise, Theory Theory can be charged with being overly imperialistic, leaving no space for other theories of the domain, or being vacuous. The debate has also brought up some important questions about the nature of Theory Theory that need to be addressed before progress can be made in the area.

The thesis is structured as follows. Chapter 1 is introductory. It introduces the notion of folk psychology - our practice of attributing psychological properties to each other. We think of ourselves and others as thinking, wanting, desiring, intending, acting, and so on. A question then arises concerning how we do this. What is causally efficacious in the production of these attributions? Accounts that explain this are called "internal accounts". This is opposed to systematisations of folk psychology that do not aim at describing how we actually attribute psychological properties. Such accounts are "external accounts" of folk psychology. The thesis is only concerned with internal accounts. I then present the two prevalent internal accounts of folk psychology: the Theory Theory and the Simulation Theory. According to the Theory Theory, it is knowledge of a folk psychological theory that is causally efficacious in the production of our psychological attributions. According to the Simulation Theory, it is our ability to imaginatively identify with others that forms the basis of folk psychology. Having this contrast in mind is invaluable for formulating the basic commitments of the Theory Theory. Chapters 2-5 are dedicated to disagreements or misunderstanding having arisen in the course the Theory Theory versus Simulation Theory debate about the nature of folk psychology. I address four major issues: the theoriticity of folk psychological knowledge, Theory Theory's alleged adherence to functionalism and denial of us having a distinctive knowledge of our own minds, and the nature of knowledge of folk psychological theory.

In chapters 2 & 3, I address the question: what does it mean to say that we have knowledge of a folk psychological *theory*? In chapter 2, I consider Stich & Nichols' proposal that folk psychological theory is a folk theory and folk theories are theories because they are bodies of knowledge. The guiding idea is that all bodies of knowledge are theories. This formulation, however, might lead to a collapse of the Theory Theory versus Simulation Theory debate since one can understand simulation as drawing on a body of knowledge. A collapse can be averted, however, if one accepts what I call the "minimal distinction". The minimal distinction is a distinction between Theory Theory and Simulation Theory in terms of the complexity of the representations that are involved in folk psychological reasoning on either account. On the Theory Theory account, these representations are always more complex than they are on the simulationist account in any given case. A problem is that there is a version of simulationism that is classified neither as a Simulation Theory, nor as a Theory Theory on the minimal distinction. But I argue that rather than that being a disadvantage of the distinction, it is a virtue. This is due to the fact that this variation of simulationism wavers between the two positions and is, if anything, best seen as a Theory Theory. Ultimately, I conclude that the Stich & Nichols' construal of the theoreticity of Theory Theory is unsatisfactory. By becoming synonymous with 'body of knowledge', the term 'theory' becomes quite uninformative when predicated of anything. Furthermore, it fails to provide a distinction between internal accounts of folk psychology that claim that although some body of knowledge is causally efficacious in the production of psychological attributions, it is not knowledge of a *theory*, and accounts that are dead serious about folk psychological knowledge being knowledge of a *theory*. A tighter notion of 'theory' leading to a somewhat stricter view of what counts as a Theory Theory, will allow us to distinguish much better between the various accounts in this area. Hence, a tighter notion should be adopted, both because it is more informative and because it carves up the domain better - it highlights important differences.

In chapter 3, I consider the more substantial reading of 'theory' in terms of 'scientific theory'. The idea is not that folk psychological theory is a scientific theory, but that it is a theory because it has important similarities with scientific theories. It is not the scientificness of scientific theories that is at issue, but the theoreticity of them. There are two prevalent ways of modelling this form of theoreticity. The one I call 'traditional', the other I call the "framework theory". The most prominent proponents of the traditional approach are Gopnik, Meltzoff, and Wellman. Their idea of theoreticity is derived from such philosophers of science as Hempel and Nagel (hence traditional). They advocate quite a stringent notion of 'scientific theory'. In fact, it is so stringent that it excludes many bodies of knowledge that we generally assume are scientific theories from being so. Therefore, I propound a weaker, but still substantial notion of 'theory' that not only not excludes theories that we generally take to be scientific theories, but that also allows a number of bodies of knowledge not seen to be *scientific* as being theories. According to this model, a theory is composed of lawlike generalisations that contain interrelated terms that refer to entities and properties that explain the data, for example by being causally efficacious in the production of them, or by being related to them in a lawlike manner. The body of folk psychological information will count as a theory on this view. 1.

Lastly, I consider the framework theory approach advocated by Carey and Wellman. It is inspired by Kuhn's theory of the nature of scientific theories. The problem is to extract from an account like Wellman's a general notion of 'theory'. It seems that either framework theory comes out as a special kind of theory, or it comes out as a number of tacit assumptions about how one constructs and applies theories. In the one case, the result seems too strong for Theory Theory. We are not, at this point, interested in showing that folk psychological theory is a special kind of theory. This might commit us to something too strong. In the second case, it seems that folk psychological theory really isn't a theory after all. Therefore, I argue that we should stay with a modified version of the traditional approach, and reject a framework theory approach.

In chapter 4, I turn to the oft quoted correlation between Theory Theory and functionalism. Various people appear of the opinion that Theory Theory is a functionalist theory and, consequently, that it is committed to a particular view of how we attribute psychological states to ourselves that doesn't lend itself to us having any kind of distinctive knowledge of our own psychological states. What has most often been supposed, is that Theory Theory is committed to semantic functionalism and metaphysical common sense functionalism. I argue that Theory Theory is committed to neither. Theory Theory is primarily an empirical theory about what puts us in a position to attribute psychological states. How its terms are defined and the metaphysical nature of the states it refers to, are not directly relevant to it.

? But we might reasonably ask about those issues.

Rejecting that Theory Theory is a functionalist theory is not sufficient to show that it is not committed to a view of self-attribution that is third personal, and that leads to a rejection of the idea that we have distinctive knowledge of our own minds. For all I have said so far, it seems that psychological attributions are based in just the same way whether they are third or first personal. This would be a symmetric view of psychological attribution. But symmetrical accounts are counterintuitive - our self-attributions seem to be based on some direct and immediate access to our own psychological states. This is an asymmetric position. I go on to argue that neither position is satisfactory. Asymmetric accounts are apt to lead to solipsism. Given that I attribute psychological states to you on the basis of observing your behaviour, and I attribute psychological states to myself on the basis of what is given to me in introspection, how do I know that I attribute the same kinds of states in the two cases? So, whereas a satisfactory account of self-attribution must allow some asymmetry between first and third personal attribution, it must not advocate complete asymmetry.

In the psychological literature, there is evidence that we have less direct access to our psychological states than we normally think we do. In particular, there is evidence that we have no direct access to the intentionality and the causal relations of our psychological states. The knowledge that we have of these aspects is inferential, hence not distinctively first personal. I then present an ontogenetic myth to pave the way for a Theory Theory account of self-attribution and self-knowledge. According to this account, something is indeed given to us in introspection - some state of affairs is presented to us. On the basis of this and interaction with other people and our environment, we come to conceptualise what is so given in a particular fashion - as folk psychological states. Hence it is an account that has both symmetric and asymmetric parts to it. Lastly, I show that this account is compatible with, or quite similar to, a number of accounts of self-knowledge that allow we have distinctive knowledge of our own minds.

Lastly, in chapter 5, I examine the thesis that our knowledge of folk psychological theory is tacit. Prototypical examples of tacit knowledge are knowledge of transformational grammar and knowledge of visual parameters. In order to evaluate the suggestion, I consider three theories of tacit knowledge prevalent in the philosophical literature. First of all, it is necessary that these three accounts accord with knowledge of grammar. I give an example of such knowledge - knowledge of *wh*-traces. Secondly, we must look at some specific examples of folk psychological knowledge to see whether it fits any of the accounts of tacit knowledge. I take a generalisation from chapter 1 as providing a prototypical example of folk psychological knowledge, but any generalisation discussed there would have done equally well. Although it is not certain that the theory theorists that assume folk psychological knowledge is tacit will agree with my formulation of the theory in chapter 1, the fact that they have provided no alternative account that can be evaluated, forces me to stay with that scenario.

Chomsky's account of tacit knowledge is relatively simple. Tacit knowledge is representational, causally efficacious in the production of thought and behaviour, and its causal powers are related to its representational content. It is, however, unconscious. One might mean different things by 'unconscious' as is clear in Stich's account of tacit knowledge. According to Stich, tacit knowledge is consciously inaccessible and inferentially encapsulated - the opposites of both are conscious characteristics. We do not have a characteristic conscious experience when presented with the content of a subdoxastic state that we have. Furthermore, the information contained in such states only have limited interaction with other psychological states. We cannot, for example, retrieve such information at will, give words to it, and so on. Considering this, it is most likely that Chomsky means 'consciously inaccessible' by 'unconscious'. Davies, for his part, suggests that the true distinction between ordinary and tacit knowledge is to be found in the Generality Constraint. The former is subject to it, whereas the latter isn't. In order that people have beliefs, they must exercise the concepts involved in the content of the beliefs. This is not the case with subdoxastic states.

I go on to consider whether folk psychological knowledge is tacit on any of these accounts. I conclude that it isn't. We assent when we are asked whether we believe that a folk psychological generalisation is true and we will do so because we have a characteristic conscious experience. We also often volunteer such generalisations. So, the knowledge is not consciously inaccessible. Secondly, our beliefs affect our folk psychological knowledge and vice versa. There seems to be little restriction on how folk psychological information combines with other information that we possess (apart from that which is tacitly known, of course), hence it is informationally integrated. Thirdly, it is conceptual in so much as it passes the Generality Constraint. Furthermore, considering recent research in experimental psychology, it seems unlikely that any account of tacit knowledge will come to classify folk psychological knowledge as tacit. Hence, I conclude that folk psychological knowledge isn't tacit.

In chapter 6, I conclude my findings. I also look at some other important issues that need to be addressed by Theory Theory, but that I do not have time to consider in detail here. They concern the acquisition of folk psychological theory, *ceteris paribus* clauses, and the compilation of a properly explanatory folk psychological theory. I give some hints at the direction in which I believe the right solutions lie.

In my research, I have benefited greatly from discussions with a number of people. I would like to thank my primary and secondary supervisors, Tim Crane and Michael Martin. Both have provided

invaluable help in the genesis of this work. My husband, Gabriel Segal, has contributed through enlightening discussions and much support. I am deeply indebted to Daniel Friesner for discussions about the nature of science and cognitive development. I have also benefited from comments by Milena Nuti and Isabella Muzio. Lastly, I would like to thank my moral supporters: Anja Matwijkiw, Matthew Patrick, and Eleni Vambouli.

London, April 2000

Chapter 1

Two Internal Accounts

Phoebe: "Oh my God. He wants me to come over and feel his bicep and more...

Rachel: "Are you kidding? I can't believe he would that to Mon... Wooh... [turns around to Joey] Joey, do they know that we know?"

Joey: "No..."

Rachel: "Joey"

Joey: "They know you know."

Rachel: "Ooh, I knew it. I cannot believe those two!"

Phoebe: "They thought they could mess with us. They're trying to mess with us. They don't know that we know they know we know. And Joey, you can't say anything!"

Joey: "Couldn't if I wanted to."

Friends, Series 5, Episode 14

Folk psychology is a practice that we are continuously engaged in in our everyday life. We attribute psychological properties to people; to ourselves as well as to others. These attributions are manifested through speech, thought, involuntary behaviour and action. I tell someone about how another acted towards me, I think that my friend is upset, I find myself blushing being complimented, and I hang up on someone who insults me. The two last situations are distinguished by the first being involuntary and the second voluntary. Psychological categories include action (walking somebody home, buying a pint, throwing oneself in someone's arms), certain kinds of involuntary behaviour (crying, blushing, shrieking, frowning, smiling, seeing and feeling something), thought (believing that *p*, supposing that *p*, phantasizing that *p*, imagining that *p*, knowing that *p*), desire (wanting *p* to be the case, desiring *x*), intention (intending to ϕ, deciding to ϕ), and emotion (hoping that *p*, fearing that *p*, loving *x*, hating *x*, being happy that *p*, being sad that *p*).

Whereas involuntary behaviour comes in many varieties, only some attributions of such behaviour count as psychological attributions.[1] The rule is that only the kind of behaviour that can be explained by reference to an agent's psychological states - her thoughts, desires, intentions, or emotions - is subsumed under psychological categories. This is the very same reason that voluntary behaviour - mostly known as action - is captured by psychological categories. The difference between action and involuntary behaviour is that whereas the former is voluntary, the latter is not. One way to flesh out this notion is to say that behaviour is voluntary if the agent could have done otherwise. Another way to look at it, is to claim that behaviour is voluntary because it springs from a decision to act, and that decision need not have been made. This, however, is more controversial (cf. Pink, 1997). Here is not the place to enter on the free will debate, so I leave the notion of the voluntary relatively unexplained, relying on our common sense notion thereof.

[1] When I talk of attributions of behaviour, I mean attributions of particular kinds of behaviour, for example laughing, stroking, and so on. It is behaviour under a particular description that is at issue.

So, both voluntary and some involuntary behaviour is explicable in terms of psychological states. Shrieking counts as a psychological category because you either do it voluntarily or you do it because you were suddenly frightened, surprised, or feeling pain. All these descriptions are folk psychological.

Neurotic and psychotic behaviours are also subsumed under psychological categories. Coming across these psychological categories, we may want to sharpen our above description and require that what counts as forming part of our folk psychological practice is not any old attribution of psychological properties, but attribution of *folk* psychological properties. Arguably, the psychological categories that subsume neurotic and psychotic behaviours, are not folk psychological categories as such. Terms like 'depression', 'schizophrenia', 'psychopathy', and 'mania' were introduced and are used by experts, be they psycho-analysts, psychiatrists, or psychologists. However, I don't happen to think that there is a sharp distinction between such psychological categories and folk psychological ones.

There are obviously folk psychological categories that are what one might call *traditional* folk psychological categories. The examples of the first paragraph are of this kind. Psychological properties like those have been attributed to people for thousands of years. My point is simply that any kind of psychological knowledge attained by specialists can be acquired by non-specialists. Thus acquired, it can come to form part of what psychological states people attribute to others and themselves outside any psychological-professional context. For example, many people now attribute others - and themselves - unconscious psychological states taken from psycho-analysis. Indeed, it has become a veritable vogue in many American movies and television series. Most of us remember memorable lines from Woody Allen movies, but now psycho-analytic discourse features large in teenage soap series. In *Buffy the Vampire Slayer*, Cordelia, the dumb beauty queen, rhetorically asks the slayer: "what is your childhood trauma?". Childhood traumas have begun to play a role as everyday explanatory constructs, sometimes with severe negative undertones. Someone who acts in weird, unpredictable, and often violent ways may be said to have one. Most people, in the western world at least, will have a pretty good idea of what that means. We might, then, want to include traumas, complexes, unconscious phantasies, and so on, among our folk psychological categories[2].

It seems that every kind of involuntary behaviour that gets to be subsumed under a psychological category has a voluntary counterpart. I can cry, blush, shriek, etc. voluntarily, or I can suppress my urge to ? cry, blush, shriek, and so on. However the behaviour is elicited, it is correct to describe it as crying, blushing, or shrieking. Nevertheless, there can be a big difference in how the more general behaviour that the agent is engaged in is described. 'He pretended to be upset' would be the voluntary counterpart of 'He was upset'. That is, someone who cries because they have *decided* to do so is pretending to be upset. It is generally assumed that in order to count as being genuinely upset, crying must be an involuntary behaviour. However, both ascriptions are folk psychological, it is just that the reasons for which someone does something feed back into the description of what they do.

1. *The Function of Folk Psychology*

Folk psychology serves at least three different functions. It enables us to:

> i. *explain* why someone does[3], feels, desires, intends or thinks as she does, ii. *predict* what someone will do, feel, want, intend or think, and iii. *understand* what someone is doing, thinking, wanting, intending or feeling.

attribution ?

I have included iii. because I don't believe that all psychological understanding can be understood as psychological explanation. Sometimes I just want to understand how someone feels, not necessarily why they feel as they do. However, at times I shall speak of

[2] This does not involve us accepting most psycho-analytic theories unquestioningly. For example, one might doubt Freud's theory of the Oedipus complex (Freud, 1900/1953), whilst maintaining that we do have a rich unconscious phantasy life that feeds into our behaviour. Accepting psycho-analytical states presumably *does* mean that we have an unconscious much like Freud conceived of it, populated by representational states that are processed at the personal level (not at the sub-personal level like tacit knowledge states, cf. chapter 5), and that play a causal role in the production of our behaviour without our consent.

[3] I use 'do' here to cover both voluntary and the types of involuntary behaviour discussed above.

folk psychological understanding as short-hand for this three-fold function of folk psychology.

We might say that folk psychology is the practice of making ourselves and others comprehensible. Such understanding forms the cement of human relations. Understanding ourselves we are better able to plan future courses of action; understanding others we have a good idea of how to relate to and interact with them. Some philosophers seem to believe that without folk psychology we would have little of the complex social interaction that we do because we need to be able to foresee the actions of others in order to regulate our own (Fodor, 1987). Others think that we can at least plan interactions with others without having to predict their actions (Morton, 1996). Instead, our interactions are based on *expectations* formed *in the process of* making decisions (cf. chapter 5). However, these expectations, themselves, seem to be formed on the basis of folk psychological knowledge about the people involved.

I think it is a point well taken, that folk psychology does not need to work *via* prediction for it to serve as an important social tool for human interaction. There is no doubt that it *does* work in this fashion sometimes, when we scheme things, for example (whether the prediction precedes, or takes place in the course of, decision making is not important here). On the other hand, much social cooperation is based on expectations, some prior to the decision making we other formed ? during it. A general expectation is that if I am polite to people, they will be polite also. It would be folly to suppose that every time we interact with people, we try to predict whether they will respond to our politeness, for example. However, I might form an expectation that a particular person will be polite during my decision making; I start out on the assumption that I will be polite, consequently that she will be polite, and then decide what else to do on that basis. Folk psychology generates expectations. Whether or not folk psychology enables cooperation *via* expectations or predictions, it enables us to live relatively stable social lives.

Being able to foresee people's reactions to one's actions is very important for future purposes. If one gets blacklisted, one might have quite a difficult time. This is why questions such as "what will she do if I do this?", "what will people think, if I do that?", and so on, are crucial in order to decide what to do. Decision making need not involve explicit answers to such questions, for often we just stick to social mores for a rule of thumb about what counts as acceptable behaviour. Nevertheless, at some level it is very important that people are relatively predictable. For our purposes as well as their own. This is consonant with the above, for my expectations are only reasonable if I assume that people are relatively predictable. This is not to say that I need to be able to predict exactly what they will do or not do, but I must have some rough picture. This is why we are very reluctant to engage in certain transactions with strangers. We cannot rule out that they will not act to our detriment. We need to assure ourselves that we don't deal with potential rapists, murderers, thieves, betrayers, and so on. Less dramatically, we need to be able to gauge people's reactions, so that we don't exclude ourselves from future cooperation with them through offending behaviour. From their point of view, it is equally important that they are predictable, in the sense just elicited, so they can count on being cooperated with. That people should be relatively understandable is crucial for social cooperation; that they should not be completely so is equally crucial. Otherwise, we should die of boredom. Predictability has its time and place.

Whatever social and emotional function folk psychology may serve, we don't tend to be instrumentalist about it when we apply it to people and other highly rational beings. We don't attribute psychological states simply because it serves certain purposes; we believe that people are in the states that we attribute them. Another way to put the same point is to say that we are intentional realists. To the untutored mind, at least, people do act, have beliefs, desires, hopes, and fears, and so on. Some people who have thought a lot about these matters come to believe that our folk psychological framework simply cannot be true (Churchland, 1981; Dennett, 1987; Stich, 1983). On the other hand, a lot of people who have also thought a lot about the matter, think that the folk psychological categories that we attribute, are largely true of us (Armstrong, 1994; Fodor, 1987; Goldman, 1993; Horgan & Woodward, 1990). Among the folk, however, there is generally little disagreement about the reality of psychological states and properties.

In what follows, I shall speak as a naïve folk psychologist. I will not address the question of whether anything actually corresponds to the psychological properties that we talk about and attribute to each other. What I assume is that we have a practice of attributing folk psychological categories to ourselves and others, and that this dictates how we conceive of each other and ourselves. I am speaking from inside this conception. It remains an open question whether one thinks it is possible to describe our folk psychological practice without being committed to it being largely true. 'Conceive of', 'understand', and so on are all intentional terms. Using such terms seems to commit one to the truth of folk psychology. However, given how many eliminative materialists that are happy to describe folk psychology - for example Paul Churchland and Steven Stich (Churchland, 1981; Stich, 1996) - we can assume that at least some philosophers think that some non-intentional reformulation of such a description is possible.

2. How Folk Psychology Works

Folk psychological predictions and explanations are only possible once there is a prior folk psychological attribution to work on. For example, we must be able to see behaviour as a particular kind of behaviour. On the other hand, once we have got a folk psychological attribution, we can do a lot with it. Take the following example:

John is afraid that it will rain

An obvious question is why John feels this way. In the absence of any further information, we can generate innumerable explanations. Some, no doubt, more farfetched than others. But normally the circumstances surrounding John's fear will give us a good idea of why he is afraid that it will rain. Let us imagine that we find out that John is having a garden party. Immediately, we are able to come up with an explanation of his feelings. If it begins to rain the party cannot be in the garden, so the garden party will be ruined. In fact, it will cease to be a garden party, strictly speaking. He really wants himself and others to have a good time, but if the garden party is ruined, this is unlikely to be the case. This partly explains why he fears that it will rain.

But the above is not enough to explain John's fear, for fear is an emotion that requires that the subject thinks that some states of affairs may come to pass that she desires not to occur. Therefore, we need to point out, not just why it is that John desires that it does not rain, but also that he thinks it might. This, then, will count as an explanation of his fear.

The next question, then, is: why does he believe it might rain? Why, indeed. Perhaps he heard the weather forecast and rain was predicted. Maybe he has seen that the sky has clouded over. Or perhaps he is of a particularly pessimistic mindset and thinks that it will probably rain since this will ruin his party and make him feel dreadful. If we know more about John, we should be able to rule out some of these possibilities. For instance, if we know he is generally an optimist, we will go for either the weather forecast or the overclouding option.

The next thing one might consider is what John will do, given that he fears it will rain. If he is particularly neurotic, we might predict that he will pace around the house, groan, wave his arms around maniacally, and so on. Or, perhaps John is a very practical person. In this case, we will expect him to arrange his house such that it will accommodate the number of guests that he is expecting. He might remove the table cloths from the tables in the garden, move some chairs inside, and so on.

But why is John having a garden party in the first place? After all, having a garden party at any time of the year is very risky if one lives in England. Is John a hopeless optimist or is he simply prey to an overly optimistic culture?

What will John do if it actually starts raining? Will he try to cancel the party? Or what if the weather clears up? Will he fall down on his knees and thank the Lord? All of these questions, and many more, can only be answered by knowing more about John. And maybe some can't be answered at all, since we just might be at loss as to how ?
to figure out how the facts, as we know them, will bear on any particular contingency. (This is not to be confused with the claim that ok
there is no answer to these questions.)

Once we start attributing psychological properties to people, we can go on for quite a bit, looking ever further back for explanations for the particular situation at hand, or predicting what they will do, think, feel, and so on, in the future. Naturally, if our knowledge of John is limited to the immediate circumstances surrounding the garden party, we won't get very far. But even if we have known John all our lives, there are definite limits to how far back we can stretch our explanations and how far forward our predictions, if these are meant to be taken in earnest.

3. *The Structure of Folk Psychology*

cf "folk psychology is a practice" — seem to have drifted away from this def^n.

In folk psychology, it is assumed that behaviour of a certain kind, psychological states, and the environment interact with each other. What the environment is like affects what we think, want, and feel. What we think, want and feel influences what else we think, want and feel, and what we intend and how we behave. And how we end up behaving affects our environment, and the whole cycle starts all over again. We may illustrate this with John above.

John believes that he is having a garden party, and he wants it to be a good garden party. A couple of hours before the party, it becomes overcast. John looks at the sky and comes to believe that it is overcast. If it is overcast, there is more chance of it raining. John comes to believe that it is likely to rain. John also believes that if it rains, it will ruin his garden party. Therefore, he comes to *fear* that it will rain. Presumably, John has some idea of what to do if it rains. He knows he can't call off the party at such short notice, so there are few options but to try to go through with the party, but inside. So John decides to have the party inside. Since he knows that he will need some tables and chairs inside, and that some things that are currently in the garden will be soaked or ruined should it rain, he starts moving some things inside. Once he has done so, he might come to believe that he has now done the best he can for a successful party. And on and on we can go.

Folk psychological understanding has two salient characteristics: i. it outlines causal relations between the explanans and the explanandum, or the prediction and the predictive basis, and ii. it makes one of these factors rational or intelligible in the light of the other. So, to return to John, the desire to throw a good garden party and the belief that it will rain cause John to fear that it will rain. But his fear also makes sense, or is rational, given his belief and his desire. As Jerry Fodor has argued, this double function of folk psychological explanations and predictions is best conceived of in the following way. The intentional or semantic relations between psychological states typically respect the causal powers of these states (Fodor, 1987, p. 13).

When we say that it is rational or that it makes sense for John to be afraid that it will rain, it is because garden parties and rain don't go together (although unfortunately they frequently occur together). Having a party is normally connected with wanting to have a good party. Given the possibility of rain, John's fear is perfectly rational. (John might end up having a good party anyway, of course). This rather lengthy link of semantic associations forms the basis of the so-called rational relation that holds between the explanans and the explanandum in a folk psychological explanation. It makes sense of the former by relating it to the latter. The same sort of link holds in other forms of folk psychological understanding. Despite the fact that it can be rather cumbersome to make this link explicit, it is one we make instantly. It should be noted that the link keeps in place the rational relation between John's desire and belief on the one hand and his fear on the other both for John himself - this is the reason that he fears - and for the observer. The desire and the belief cause the fear *and* rationalise it.[4]

What is meant by 'rational' is, of course, very broad. It is not restricted to some narrow sense such as 'logical'. For it applies not only to transitions among beliefs, but also to decision making, action, and emotion. It may be stretching the notion 'rational' a bit far at times - in which sense is it rational for Claire to cry because Paul says she's overweight and ugly, for example? In these cases, perhaps, it is better to use the expression 'makes sense'. Whether there is any non-circular way of fleshing out this notion is not clear at present. But for our purposes, we can simply note that folk psychological understanding rationalises thought, behaviour, and so on, in the sense of 'makes sense of'.

[4] Whereas certainly all participants in the Theory Theory versus Simulation Theory debate accept this construal of folk psychology, not all philosophers have been happy to do so. G. E. M. Anscombe and A. I. Melden are cases in point (Anscombe, 1957; Melden, 1960). They thought that reason explanations and causal explanations were necessarily distinct kinds of explanations and mutually exclusive. Melden maintained that giving a reason for an action is simply another way of describing the action. I will not go into the details of his argument here. I take it that Donald Davidson has shown that the main argument in favour of this view does not work, and thus that reasons *can* be causes (Davidson, 1963).

Many folk psychological states are also known as propositional attitude states. Propositional attitude states are so called because they are attitude states with a propositional content; for example, the belief that the moon is full, the desire that an enemy will come to a particularly nasty end, and the hope that the meaning of life will soon be revealed. The belief, the desire, and the hope are all examples of attitudes; *the moon is full, an enemy will come to a particularly nasty end*, and *the meaning of life will soon be revealed*, are propositions. Propositions present a particular state of affairs, they present the world as being in one way or another.

However, I prefer to look at psychological states in terms of Why? mental representations. A painting, a photograph, a word, and a symbol are all representations. A painting by Constable, say, will typically represent some lush landscape or other. A photograph represents what it is a photograph of. The word 'word' represents words, and so on. Psychological states are relations to representations in our minds. They have a psychological mode (cf. attitude) and a psychological content (cf. proposition). Psychological content represents something or other, for example that the moon is full. Another way of saying the same thing, is to say that psychological states are about something, that they are intentional. A representation need not represent a state of affairs, a situation. A representation can represent a thing. Therefore, not all mental representations need be propositional, although all propositional attitudes are relations to mental representations. Lastly, mental representations can represent non-existent objects and states of affairs, for example *Santa Claus* or *the present Kind of France is bald.*

Most, but not all, psychological states are representational. It is just possible that all folk psychological states are representational, but disagreement reigns. Someone like Searle believes that they are not (Searle, 1983). States such as pain, anxiety, depression, elation, and melancholy are psychological but not intentional. How could they be? What would be their content? Tim Crane, on the other hand, argues that states such as these *are* representational (Crane, 1998). They just represent something in a slightly different way - in a broader sense, perhaps - than the more traditional psychological states, like belief. For example, pain is directed at a physiological event in some part of my body, depression represents the world as being "a pointless and colourless place: nothing seems worth doing" (p. 242).

Another, closely related, issue is whether all folk psychological states are propositional attitude states. Some say they are not (Crane, 1995 *and* Searle, 1983). I may love a cat, or hate spiders, neither of which is easily put in propositional terms. On the other hand, there is disagreement about whether desire is a propositional attitude. Some say that 'I desire a cup of coffee' really captures the propositional attitude in which I desire it to be the case that I have a cup of coffee (Crane, 1995, p. 26). Others, such as Michael Martin, maintain that this is unreasonable for certain desires because of their temporal aspect (Martin, MS).

tell us

4. Referential Opacity

The truth of a psychological ascription rests on the subject standing in the relevant relation to the proposition or term that is the content of the psychological state that she is ascribed. For example,

(a) Samantha believes that Isak Dinesen wrote *Seven Gothic Tales*

is true if and only if Samantha believes that Isak Dinesen wrote *Seven Gothic Tales*. Whether or not Isak Dinesen indeed wrote *Seven Gothic Tales* is irrelevant to (a). Of course, the truth of the assertion:

(b) Isak Dinesen wrote *Seven Gothic Tales*

does depend on Isak Dinesen having written *Seven Gothic Tales*. And for Samantha's belief to be a *true* belief, (b) must be true. But this is different from it being true that Samantha has the relevant belief.

Sentences expressing psychological states have a peculiar characteristic that is indicative of the nature of a psychological state; the content clause is opaque in the following sense. Under normal circumstances one can intersubstitute co-referring terms in a sentence whilst keeping its truth value constant. For example, Isak Dinesen was a pen name used by Karen Blixen, so

(c) Karen Blixen = Isak Dinesen

This means that we can substitute 'Karen Blixen' for 'Isak Dinesen' in (b), without changing its truth-value, creating

(d) Karen Blixen wrote *Seven Gothic Tales'*

(d) is true if and only if (b) is true. However,

(e) Samantha believes that Karen Blixen wrote *Seven Gothic Tales*

can be true even when (a) is false. And (a) can be true consonant with (e) being false. Samantha might know (b) without knowing (d), because she might not know (c). This phenomenon is also known as referential opacity or the intensionality of propositional attitude ascriptions. Co-referring terms cannot be intersubstituted *salva veritate* in the content clause. We can, however, substitute outside it. So, if Samantha is the girl who won the lottery, we can make the attribution:

(f) The girl who won the lottery believes that Isak Dinesen wrote *Seven Gothic Tales*

The last thing I want to mention is how psychological understanding varies across the population. Some people are very good at understanding others, and others are very bad indeed. In general, women are better than men (Baron-Cohen, O'Riordan, Jones, Stone & Plaistead, 1999; Baron-Cohen, Jolliffe, Mortimer & Robertson, 1997). Children with older siblings are better than only children (Perner, Ruffman & Leekam, 1994). Teenagers are better than children, adults better than teenagers, and so on. People with autism or Asperger Syndrome are extremely bad at it, although they may eventually come to be tolerably good at it (Baron-Cohen, 1995; Frith, 1989). Practically all humans engage in folk psychology - to a larger or smaller extent.

5. *Internal and External Accounts*

There are two kinds of accounts that one may give of our ability to engage in folk psychology (cf. Stich & Ravenscroft, 1996). One might provide a theory that accounts for folk psychology, but that is not committed to mapping the psychological events that are causally efficacious in the production of folk psychological attributions. This would be an external account of folk psychology. An internal account, on the other hand, provides a description of the causally efficacious mechanism.

The distinction between external and internal accounts is, perhaps, best illustrated by a comparison to grammar. We all learn external accounts of grammar in school whether for the purpose of becoming increasingly aware of our mother tongue or learning a new language. These grammars posit a number of rules or principles that are sufficient for generating the relevant syntax. However, if Noam Chomsky is right, these grammars do not correctly describe the principles that are causally efficacious in the production of the utterances of native speakers. Chomsky's is an internal account of grammar. He is not concerned with useful systematisations, but with capturing what produces judgements of grammaticality (Chomsky, 1975, 1986).

Sometimes 'folk psychology' is used to refer to some body or systematisation of psychological intuitions of the folk. Such an account of folk psychology (in my sense of the term) need only be external. I will only concern myself with internal accounts of folk psychology here.

6. *Knowledge of Folk Psychology*

There are a number of internal accounts of folk psychology. The two most prevalent are the Theory Theory and the Simulation Theory. The main focus of this thesis is the Theory Theory and, to the extent that it has engendered fruitful debate about the nature of Theory Theory, the Simulation Theory. One may characterise the Theory Theory versus Simulation Theory debate in terms of knowledge. It would be natural to understand the Theory Theory as maintaining that a proper account of folk psychology must be in terms of knowledge-that, whilst the Simulation Theory propounds an account of folk psychology in terms of knowledge-how. → yields K. that

Knowledge-that is propositional knowledge. Knowledge of theories, or parts of theories, is propositional knowledge; knowledge that E=mc^2, or that jackals mate for life, for example. Knowledge-how is, unsurprisingly, non-propositional. It is often regarded as an ability or a skill, such as knowing how to ride a bike, how to swim, how to combine colours, and so on.

Not every ability needs to be based on know-how. Alternatively one might say that not everything that appears to be an ability on the face of it, is an ability. eg? However, I prefer the first way of talking, since that does not prejudge too many issues. Cognitive science has the, some think nasty, habit of explaining abilities in terms of knowledge-that. Vision research and transformational grammar are good examples of this (Chomsky, 1975; Marr, 1982). Forming three dimensional images from two dimensional retinal stimuli, or knowing how to produce and comprehend utterances are very much examples of what one might call abilities. Yet they seem fruitfully explained in terms of know-that. There may even be people who believe that all abilities can be explained in terms of know-that. However, some people are very sceptical about the know-that tendencies of the cognitive science community - at least of classical AI (for example Heal, 1994b).

There may even be abilities that one can divide into a know-that and a know-how part. Knowledge of theories, for example, appears to be like this. For example, one might know a scientific theory and not know how to apply it. This is why science books are full of exercises, encouraging you to *use* the knowledge that you have acquired. Knowing formulas, results of experiments, and the like (know-that), does not exhaust a scientist's knowledge. She also knows how to device experiments, when the laws she knows are applicable, and so on (know-how). Thus, not all areas of cognition or ability can be analysed exclusively in terms of know-how or know-that.

What started out as an adamant debate to prove the other wrong, has now become a more laid back, having most simulationists admitting that there is some know-that in simulation and most theory theorists admitting that there is some know-how in our folk psychological attributions. They might not quite put it this way. I have done so in order to promote a first understanding of the issues. As we proceed, the distinction between the two theories will become sharpened.

7. An Outline of the Theory Theory

According to Theory Theory, knowledge of a folk psychological theory is causally efficacious in the production of our folk psychological attributions. The first 'theory' in Theory Theory refers to this theory. There is a second 'theory' because Theory Theory is, itself, a theory, that is a theory about folk psychology. In short, the Theory Theory is a theory about folk psychology that maintains that knowledge of a folk psychological theory is causally efficacious in the production of (folk) psychological attributions.

A theory contains a number of statements. Knowledge of such statements is productive of folk psychological attributions. Probably the best known statement of folk psychological theory is:

> (G1) If *a* desires that *q* and believes that if *p*, then *q*, then *a* will attempt to bring it about that *p*, *ceteris paribus*

I shall sometimes refer to this as the "action generalisation". However, there is one important problem with (G1). It seems simply false as it stands. For example, I want to feel really great about myself, and I believe that if I sniff cocaine, then I will feel really great about myself. Yet I do not try to bring it about that I sniff cocaine. Or, I want to lead my life like a Proust or Ruskin, travelling, speculating, writing at my own ease without having to worry about earning a living. I believe that if I had been born to very rich parents I could have done just that. I do, of course, not try to bring it about that I was of rich parents. Not just because I cannot do so, but because I don't believe that I can do so. So, a reformulation of the idea underlying (G1) is in place. Churchland has provided the following pretty exhaustive formulation: (Churchland, 1970, p. 221)[5]

CP clause

[5] I formulate this as a generalisation, not as a law as Churchland does. In chapter 3 I shall consider the lawlikeness of folk psychological generalisations.

> (G1)* **If** (1) *X* wants to Ø, and
>
> (2) *X* believes that *A*-ing is a way for him to bring about Ø under those circumstances, and
>
> (3) there is no action believed by *X* to be a way for him to bring about Ø, under the circumstances, which *X* judges to be as preferable to him as, or more preferable to him than, *A*-ing, and
>
> (4) *X* has no other want (or set of them) which, under the circumstances, overrides his want Ø, and
>
> (5) *X* knows how to *A*, and
>
> (6) *X* is able to *A*,
>
> **then** (7) *X A*-s

It is, perhaps, not entirely exhaustive. One might want to add the following clause:

> (8) *X* does not believe that the outcome of *A*-ing is such as to make it impossible or too difficult to bring about *Æ*, which is something else that *X* wants as much as, or more than, Ø.

(8) stresses the fact that often when there is no direct conflict between our desires - say the desire to go on a world cruise and the desire to pay off one's mortgage - but only between the results of acting on such desires. If I go on a world cruise, I cannot pay off my mortgage, and *vice versa*. One might say that this is no fault in my desires, but due to the unfortunate way in which the world is set up. (Why, for example, can't we have our cake and eat it too?) Churchland regards at least (4) and (6) as *ceteris paribus* clauses. However, there are many ways in which one might think of a generalisation and the conditions under which it is true. One might, for example think that all of (3)-(6) and (8) are *ceteris paribus* clauses. The idea would be as follows. (7) is true if (1) and (2) are true *ceteris paribus*. However, an agent has more than just one desire and many more beliefs than is represented in (G1)*, and many more generalisations hold true of how these are related to each other. So, in actual fact, *ceteris* are rarely *paribus*. As I understand Nancy Cartwright, her ideas about laws of physics are similar (Cartwright, 1983). Quantitative laws, like that of gravity, for example, are only approximately true. At any one time a great number of forces, etc., are at play. It is only in the laboratory that the law of universal gravitation is true, because only there are there no other

forces at play (or hardly any, anyway). Now, imagine that we could isolate psychological states. If there were someone, *X*, with just the desire to *Ø*, and the belief that *A*-ing would be a way to *Ø*, then *X* would *A*. However, any real person has a great number of desires and beliefs at any one time. A great number of generalisations or laws hold true of these also. Once you factor those in, you will find that (G1*, (1) and (2)) hardly ever holds true. *Ceteris paribus* clauses plot the various factors to be taken into consideration when determining whether we can infer (7) from (1) and (2). The parallel between the case of physics and that of folk psychology is not complete, but sufficient to stress the idea that one can regard the specification of circumstances under which a law or a generalisation holds true as *ceteris paribus* clauses.

There is a choice to make of whether to regard the specification of the conditions under which a folk psychological generalisation holds true as something that is built into the generalisation itself or whether to regard it as being *ceteris paribus*. The choice might have consequences for whether one thinks that the generalisations can be regarded as laws, or whether we should think of the regularities of nature in different terms (Cartwright, 1983). However, for my purposes this is not relevant. In chapter 3, I will have more to say about the nature of folk psychological generalisations, but nothing there will determine how one must understand specifications of circumstances under which such generalisations hold true. This is an issue that might divide theory theorists. That is, which ever view one takes, one will remain a theory theorist. It is a question that can be left open in foundational work on the Theory Theory.

Int or ExtTT? Might matter.

The more specific role that knowledge of folk psychological generalisations can be put on argument form. I shall use (G1) as a shorthand for (G1)* thereby avoiding a very cumbersome formulation. If I know (G1) and I know what a person wants and what a person thinks, I can make a deduction of the following kind:

a. (G1)
a desires *q*
a believes if *p*, then *q*
a will attempt to bring it about that *p*, *ceteris paribus*

(G1) also serves well in inductive arguments. For example:

b. (G1)
a believes that if *p*, then *q*
a attempts to bring it about that *p*
a desired that *q*, *ceteris paribus*

c. (G1)
a desires that *q*
a attempts to bring it about that *p*
7. *a* believes that if *p*, then *q*, *ceteris paribus*

Neither ~~a.~~ b nor ~~b.~~ c are valid arguments, but we often make inferences of this type because they often lead to true conclusions, or because, in the lack of any further information, this is the best we can do (Kahneman & Tversky, 1982; Nisbett & Ross, 1980; Wason & Johnson-Laird, 1972; Tversky & Kahneman, 1993). Affirming the consequent is a good example of such reasoning.[6] Suppose we know that if it has been raining, the streets are wet. We go outside, see that the streets are wet, and conclude that it has been raining. This is not a valid argument, but there is a pretty good chance it is true. It is not necessarily irrational to reason fallaciously if you do so because of limited processing abilities and time (Cherniak, 1986).

But (G1) can also be used to figure out what someone believed or desired given information about what they desire or believe and what they did, given another simple generalisation:

(G2) **If** *a* attempts to bring it about that *p*, and (i) *a* has the ability bring it about that *p*, and (ii) the circumstances are such that *a* can bring it about that *p*, **then** *a* brings it about that *p*, *ceteris paribus*.

Given (G1) and (G2), we can make the following inferences:

[6] Charles Sanders Peirce called this *abduction* and maintained that it played a very important part in scientific reasoning. As a rule of inference, it works only against a background of what counts as the best possible explanation in the area (Peirce, 1933, 7.199-202). The same idea is sometimes known as "inference to the best explanation".

d. (G1) and (G2)
a believes that if *p*, then *q*
a brings it about that *p*
a desired that *q*, *ceteris paribus*

e. (G1) and (G2)
a desires that *q*
a brings it about that *p*
a believes that if *p*, then *q*, *ceteris paribus*

But for either d. or e. to be true, *ceteris* must be *paribus*. Someone like Corto Maltese might desire to find a particular treasure and find it, but by accident. For example, he stumbled into a suit of armour, it fell on the floor and broke into many separate pieces, thereby revealing the coveted treasure. In this case, we cannot infer that he believed that if he tore the suit of armour apart, he would find the treasure. In general, however, reasoning as in d. or e. is a pretty good bet.

Sometimes we have very little information about an agent to go on. Maybe we only know what a person either believes, desires, attempts to do, or does. In these cases, the action generalisation is not particularly useful. Unless, that is, I know other things about the person that will allow me to infer what other psychological states they are in. Imagine that I know that Bob wants a beer. I know that he has beer in his fridge, that he is not a teetotaller, and that he is not trying to cut down. It seems to me safe to assume that Bob believes that if he goes to the fridge and takes a beer, he will come to have a beer. I can then apply (G1) since I take it that *ceteris* are *paribus*. Hence, I can infer that he will go to the fridge and take a beer. To get this far, I used another useful generalisation:

What evidence do we have that these are the laws we use?

> (G3) **If** *p* is the case, and (i) *a* has been exposed to *p*, and (ii) *a* was paying attention when *a* was exposed to *p*, **then** *a* believes that *p*, *ceteris paribus*.

For example, I will only attribute you knowledge of some fact if it occurred in your immediate environment (i) and only if you were paying attention when it did.[7] But in the case of Bob, how do I know that he has been exposed to the fact that there is beer in his fridge? Somebody else could have put it there while he wasn't looking. In general, however:

> (G4) **If** *p* is a fact that plays an important role in *a*'s culture, environment, job, or interests, and (i) *p* is salient, and (ii) *p* is not being concealed from *a*, **then** *a* believes that *p*, *ceteris paribus*.

Given that Bob is a heavy beer drinker, we can assume that beer is an interest of Bob's, and that he will know whether or not he has beer in the fridge. Unless, of course, Bob has a wife who also likes to drink, and who wants her beer for herself. She might hide the beer in the fridge - behind all the preserves, for example.

Perhaps a more straightforward way of assuring ourselves that Bob knows that there is beer in the fridge, is that he has seen it. Here we use a generalisation of the following kind:

> (G5) **If** *a* perceives that *p*, and (i) *a* was paying attention, and (ii) *a* did not have countervailing reasons to believe that ~*p*, **then** *a* comes to believe that *p*, *ceteris paribus*.

Generally, we assume that perception is a truth requiring relation between a perceiver and the world. One cannot perceive things that are not the case. However, if we extend the use in a Cartesian way to apply to how things *seem* to us, we may say things like: 'I see that the stick is bent in water'. In this case, (G5, ii) would apply. I would not infer from this visual experience that the stick *is* bent in water.

How do I know that Bob has *seen* that there is beer in the fridge? Well, the application of (G5) builds on (G6):

> (G6) **If** *p* is detectable by normal human's sense organs and *x* occurs somewhere in *a*'s more immediate environment and
> (i) to see *x*: *a*'s eyes must be directed towards *x* and *a*'s line of sight must be unimpeded;

[7] I do not want to claim that there are never situations in which (i) to (ii) hold, but (G3) is false. In fact, I don't want to claim that about *any* of the conditions of the generalisations that I give.

(ii) to hear *x*: *x* must occur in the general vicinity of *a*, depending on its strength - if very strong *x* can be very distant from *a*, if weak *x* must be quite close to *a* ;

(iii) to touch *x*: *a* must be spatially contiguous with *x*;

(iv) to smell *x*: *x* must occur in the general vicinity of *a*, depending on its strength - if very strong *x* can be quite distant from *a*, if weak *x* must be quite close to *a* and *a*'s nose pointed in the general direction of *x*;

(v) to taste *x*: *x* must be in *a*'s mouth or touched by *a*'s tongue directly, or indirectly via lips or fingers, or the like, and

(vi) *a* pays attention to *x*,

then *a* perceives *p ceteris paribus*

Suppose I only know what someone wants. I might then apply something like (G7):

> (G7) **If** *a* wants to ϕ, and (i) *a* has no stronger desire to ψ that directly conflicts with ϕ*-ing*, and (ii) the consequences of ϕ*-ing* are not such that it excludes *a* from ψ*-ing* if ψ*-ing* is something *a* desire as much or more than ϕ*-ing*, **then** *a* will try to ϕ.

There are certain situations where it seems unnatural or even impossible to apply (G1). If I want to lift my arm, go for a run, have a nap, and so on, I just do so. In these cases, my action most naturally falls under (G7). I don't know what beliefs to appeal to explain lifting my arm, going for a run, or having a nap using (G1). Apart from these cases, we often use (G7) as a shorthand for (G1).

In all of these listed generalisations, 'desire' is not be understand in a narrow sense, most commonly connected with an intense or at least distinctive phenomenology. It should be understood more broadly as a 'pro-attitude'. A pro-attitude is a motivational state that figures in the explanation of action together with belief. When, for example, I act from a feeling of obligation, we might say that I have a pro-attitude towards the projected result of my action, but I need not *desire* that result in the above sense.[8] Similarly, there is a marked difference in

[8] I cannot here enter the debate concerning whether moral judgements, themselves, are motivating in the absence of any desire to act in accordance with them (Kant, 1785/1993; McDowell, 1978; Smith, 1994). I shall simply assume that if *a* feels under an obligation to ϕ, then *a* has a desire to ϕ.

my desire to spend the night with some fascinating creature, and my desire to pay the gas bill. One way to put that difference is to say that the first is a desire proper and the second a pro-attitude. There is another distinction one might draw between desires; that between means-desires and end-desires. There are things that I desire in themselves, and things that I desire as a means to something else that I desire:

> (G8) If a desires to ϕ, and a believes that to ϕ, a must ψ, and there is nothing else that a desires as much as to ϕ, or more, that becomes impossible or very difficult once a has ψ-ed, then a will desire to ψ, *ceteris paribus*.

I may, for example, really desire to have my hair coloured green because I want to be cool. Means-desires and end-desires are largely relative. My desire to be cool is an end-desire compared to my desire to colour my hair green. However, my desire to be cool may be a means-desire in relation to my desire to be admired. It is natural to think that, ultimately, there is only one end-desire: happiness (cf. Aquinas, 1989; Aristotle, 1976). However, as Aristotle was quick to point out, there are many ideas of what happiness is.

(G8) is very useful if one does not know someone's means-desires but only their end-desires. Frequently, one needs to know someone's means-desires to figure out what they will do, since there are often many ways of satisfying an end-desire. (G8) helps you determine someone's means-desire on the basis of knowledge of their end-desire and beliefs.

The above shouldn't lead one to expect that all folk psychological generalisations are of the very abstract nature presented above. There is going to be a host of generalisations that are much more specific, for example: (Churchland, 1988, p. 211)[9]

> (G9) A person who suffers severe bodily damage will feel pain, *ceteris paribus*.

> (G10) A person who is angry will tend to be impatient, *ceteris paribus*.

[9] Churchland, himself, does not attach *ceteris paribus* to these generalisations. This is unwise, however. An obvious counterexample to (G9) ~~and (G10)~~ is a person in a coma. I have therefore added *ceteris paribus* clauses.

Folk psychological theory also includes classification statements, such as:

(G11) All aches are pains

(G12) Some pain is emotional pain

(G13) Emotional pain is not physical pain

Before ending my survey of what folk psychological theory looks like and how it works, a line about its application is in place. How do I learn to tell that a person acts or behaves in a way that allows intentional explanation, for example? I think Theory Theory is only committed to the idea that knowing folk psychological theory is necessary for making folk psychological attributions. It need not be sufficient also. In fact, I think it would be unwise to make such a strong claim. As we have already seen, knowing a theory does not necessarily imply knowing how to use it. Knowing a scientific theory, for example, is not sufficient to knowing how to apply it. This is why teaching science always involves examples, tests, and experiments. It takes experience and skill to know how to apply a theory, and it is up to each individual to acquire it. To put this in terms used earlier, knowing how to apply a theory is, perhaps, more a matter of know-how than of know-that.

8. Knowledge of Folk Psychological Theory

When the theory theorist says that we all have knowledge of a folk psychological theory, what exactly does she mean? That is, should we regard folk psychological theory as a theory known by everybody capable of making (folk) psychological attributions? We might agree with David Lewis that only those generalisations (he says 'platitudes') that everybody knows and everybody knows that everybody knows them, should be included in folk psychological theory (Lewis, 1972). Alternatively, one might choose to say that what everybody knows, and everybody knows that everybody knows, is the *core* of folk psychological theory, allowing that some people may be more knowledgeable than others. What exactly will count as the core of folk psychological theory is an interesting question. Certainly all of (G1)-

In what sense of 'know' does everybody know G1-G8, know that ----?

(G8) are part of the core, but it seems reasonable to suppose that (G9)-(G13) also form part of it. As a rule of thumb, we can suppose that the core of folk psychology is the part of it such that, once mastered, allows for standard human interaction. which is?

I think it preferable to regard Theory Theory as maintaining that we all know at least the core of folk psychological theory, and that this knowledge is causally efficacious in the production of our (folk) psychological attributions. This explains why some people are better than others at understanding other people - they know more than the core. It is, of course, possible to explain this in terms of application - some people are much better at applying their theory than others. However, if combined with the fact that some people have extra knowledge, it becomes more plausible. I think it is difficult to explain the big differences in the population simply in terms of some being better at applying folk psychological theory than others. This is not to deny that significant differences can arise from this. Those that show great psychological understanding of others tend to spend more time thinking about people and why they do, think, and feel the way they do, they pay more attention to people, pick up on subtle signs quicker, and so on. However, this is unlikely to account for all of the differences - some of these seem to be differences in knowledge-that. For example,

performance / competence

Such as? Emp evidence?

argument?

> (G14) If *a* acts very arrogantly, *a* is either arrogant or insecure, *ceteris paribus*.

is something most of us learn with age, but many still seem to take it for granted that if someone behaves in an arrogant manner, they are arrogant. Still, there are many among us who don't yet understand this.

(G14) has a different flavour from other generalisations we have come across before. It deals with character traits. Now, character plays a large role in folk psychology. Most of us feel more confident about how a person is going to act, when we have decided what their character is like. Knowing this seems drastically to reduce the different ways in which such a person will react in specified circumstances. We say things like "she wouldn't do that, she's not dishonest", for example. Hence, we should expect there to be a sizeable number of generalisations dedicated to character. Some psychologists and philosophers think that we are mistaken in believing that there are character traits (for example, Nisbett & Ross, 1980; Harman, 1999). Many experiments have failed to unearth any correlations between character trait and behaviour (for a survey, see Ross & Nisbett, 1991). It is possible, then, that character is one issue on which folk psychological theory is mistaken. Nevertheless, it is indisputable that we do believe that people have characters. There may well be many other kinds of folk psychological generalisations. I make no claim to having unearthed them all. I have merely attempted to give an outline of the theory.

Claimed by Horgan & Woodward 1990 that the role of character in FP has changed since, say, the Regency Period.

9. The Ontogenesis of Folk Psychological Knowledge

Before we go on to the rival theory of folk psychology, the Simulation Theory, we should consider one more aspect of Theory Theory: the acquisition of folk psychological theory. Sometimes Theory Theories of development are known as Diachronic Theory Theories (Segal, 1996). What I have presented so far is a Synchronic Theory Theory. For good reasons, the developmental aspect is normally left in the hands of child psychologists. That is certainly where I shall leave it, but it is highly pertinent to look at the some of the prevalent views of how knowledge of folk psychological theory comes about. One way to look at the different Synchronic Theory Theories is to see how they answer the following three questions:

1. Is our ability to acquire folk psychological theory domain specific or domain general?
2. Is folk psychological theory innate?
3. Does the acquisition of folk psychological theory involve conceptual change?

What is meant by an ability being domain specific, is that it is dedicated a particular task, for example the production or comprehension of grammatical utterances. When an ability is domain general, it can be applied across the board, for example to draw inferences (all observed *x*'s up until now have been *B*, therefore all *x*'s are *B*). If some ability, capacity, or information is innate, it is there

Secondary sex characteristics.

from the time we are born. It is not necessarily operative from our birth, it may only start to function later in life. And it may be interfered with by an inhospitable environment. For example, our ability to reproduce is innate. However, we only start to be able to reproduce when we are between 11-14 years old. Girls who are anorexic or otherwise starved, often are incapable of reproducing themselves - an example of the effect of the environment on our innate capacities. You need to have a certain body weight in order to be fertile, hence you need to eat a certain amount of food. Conceptual change can mean a number of things. When I use 'conceptual change' here, I use it in Susan Carey's (1985) sense. Carey draws a distinction between restructuring in the weak sense, which can be regarded as enrichment of concepts much as one would expect from someone's concepts who becomes an expert in the field where the concepts apply, and restructuring in the strong sense, where there is an important change in the theory in which the concepts figure. In order that one can talk of theory change, there must be "changes in the domain of phenomena to be accounted for by the theory, changes in explanatory mechanisms, and (most importantly) changes in individual concepts" (p. 187). What is meant by "conceptual change" here is restructuring in the stronger sense. Therefore, theorists that are classified as denying that there is conceptual change should not be regarded as necessarily denying that there is conceptual restructuring in the weak sense. (For more on theories and theory change, see chapter 3.)

cf

Potentially eight different Diachronic Theory Theories are possible when judged on how one might answer 1.-3. I will only consider two such classes of Theory Theories. The first used to be known as the Child as Scientist position, but now 'the Scientist as Child' position is the proponents' preferred label for it. It is espoused most prominently by Alison Gopnik, Henry Wellman, Andrew Meltzoff, and Josef Perner (Gopnik & Wellman, 1992 and 1994; Gopnik & Meltzoff, 1997; Perner 1991). They think that our ability to acquire folk psychological theory is not innate in any profound sense, that it is domain general, and that it involves at least one important conceptual change. The second, I dub the Modularist Theory Theory. This position is advocated by Simon Baron-Cohen, Jerry Fodor, and Alan Leslie, among others (Baron-Cohen, 1995; Fodor, 1992; Leslie, 1987 and

Why? A mechanism could be domain-specific, innate & involve, in its acquisition, conceptual Δ, but not be info encap.

1994). According to it, our ability to acquire folk psychological theory is innate, domain specific, and involves no conceptual change.

According to the Scientist as Child Theory Theory as presented by Gopnik, Meltzoff, and Wellman, children are born knowing a "starting state" folk psychological theory (Gopnik & Wellman, 1994, p. 281): "We are born with certain kinds of psychological knowledge that begin a process of theory development and revision." This theory, however, is very different from a fully developed folk psychological theory. At best we can say that it is a proto-folk psychological theory. Children are also innately endowed with certain theorising abilities, the ability to reason deductively and inductively, for example. Children's proto-theory is defeasible and does, as a matter of fact, undergo important changes before it becomes recognisable as folk psychological theory. The theory is developed and ultimately changed in response to the evidence and its internal coherence. According to these psychologists, this development is importantly similar to the development of scientific theories (cf. chapter 3). Any theory goes through various stages, of enrichment, of addition of auxiliary hypotheses, of new theoretical apparatus, and finally of outright theory change.

The most studied conceptual change in a child's folk psychological theory occurs around the age of 4. At this point, children move to a representational theory of mind from what is sometimes called a "copy" theory of mind (Wellman, 1990). According to the copy theory, there is some direct relation between what is in the mind and what is in the world that excludes misrepresentation. The dividing line is generally drawn at the passing of false belief tests. False belief tasks are tasks that elicit the understanding that people children have of beliefs. A typical task is the smarties task (Gopnik & Astington, 1988). Here children are shown a smarties box and asked what's inside it. They typically reply: "smarties". The box is opened to reveal that it contains pencils. The children are then asked what they thought was in the smarties box before they opened it. Most children under the age of 4 answer: "pencils". Once children answer "smarties", they are said to have passed the false belief task. What they come to understand is that psychological states are *representations* of reality which means that they can fail to correspond to it. Notice, that it is typical of the false belief test, that children correctly attribute false beliefs to themselves at the same time they correctly attribute false beliefs to others. On tests, these abilities do not come apart. It is interesting to note that there is evidence that children come to understand the representationality of other psychological states, such as desire and intention, as well as perception, before that of belief (Astington & Gopnik, 1988; Gopnik & Meltzoff, 1997; Gopnik & Slaughter, 1991). However, it is generally assumed that only once children pass the false belief test, they can be said to possess a proper representational theory of mind. There is still a long way for children to go before they can be said to possess a full-blown folk psychological theory. Whether this is to be understood as involving conceptual restructuring in the weak or strong sense, I cannot go into here.

I take Baron-Cohen as representative of a modular theory theorist. There are, however, many modularist positions, not all of which will answer 1.-3. in the way that Baron-Cohen does. Baron-Cohen understands the child's acquisition of folk psychological theory as the result of the operation and maturation of a number of different modules: an Intentionality Detector (ID), an Eye Direction Detector (EDD), a Shared Attention Mechanism (SAM), and a Theory of Mind Mechanism (ToMM) (cf. Leslie, 1987).

The ID and the EDD serve to detect goal directed behaviour and direction of gaze, respectively. They are both operative more or less at birth. The ID allows you to identify actions and agents - although initially it may identify too much as either - in the absence of any theoretical knowledge. We are simply constructed such as to conceive of self-propelled motion as action and that which is doing the self-propelling, an agent. Thus, the ID furnishes the child with some crude idea of goals and desires - crude because not fully intentional. Unfortunately, Baron-Cohen does not give us much of an idea of what such semi-intentional states or relations are like. What seems to be at issue is some kind of non-representational directedness towards an object.

The EDD gives the child an idea about what is on a person's mind by detecting what they are looking at. For example, in certain situations one can infer what a person thinks from what they are looking at. Eventually, this will come to provide the basis of the inference 'if *a* sees that *p*, *a* will come to believe that *p*'.

Both ID and EDD use dyadic representations:

Dyadic representation: $R(a, p)$

R is a semi-intentional relation between an agent a and either an object, a state of affairs or a proposition p, all according to how conceptually sophisticated one assumes young children to be. The 'wants' (in the ID) and 'sees' (in the EDD) that are slotted into the R-position are semi-intentional precursors of the intentional states that are normally expressed by these terms. Maybe 'p is the goal of a' better expresses this semi-intentional aspect than 'a wants p' does.

From about 9 to 18 months, the SAM comes on-line. As opposed to both the ID and the EDD, the SAM creates triadic representations:

Triadic representation: $A(a, A(b, p))$

A is what Baron-Cohen calls the semi-intentional relation of 'attending to' holding between an agent a and some state of affairs, object, or proposition p.[10] This relation is most commonly 'see', but can be also be 'hear', 'touch', 'taste' or 'smell'. a and b are both agents. Either a or b must be the self, and the other another agent, such that $a \neq b$. Triadic representations not only allow the child to represent a common view on the world, but also to represent to herself her own psychological states - although her grasp of them is tenuous at this stage.[11]

[10] I have used 'A' here instead of 'R' simply to stress that the relation in question is that of attending. However *As* are a subclass of *Rs*.

[11] Baron-Cohen's position is puzzling in a number of ways. Firstly, it is unclear whether the kind of triadic representation described by him is sufficient for shared attention. It appears only to express the self (or an agent) attending to another agent's (or the self's) attending to something. This, however, is not really shared attention. Shared attention should minimally involve the self attending to something *and* to someone else's attending to that something, thus:

$$A((a, p) \wedge (a, A(b, p)))$$

But, even this way of putting shared attention leaves something out. Presumably *both* agents must do the attending that only a does above. So, we get:

$$A((a, p) \wedge (a, A(b, p))) \wedge A((b, p) \wedge (b, A(a, p)))$$

This is certainly better, but perhaps not good enough. Surely, both agents need to be aware that the other agent is attending to their attending to their attending for us to have proper joint attention. Thus:

$$A((a, p) \wedge (a, A(b, p))) \wedge A((b, p) \wedge (b, A(a, p))) \wedge A(a, A(b, A(a, p))) \wedge A(b, A(a, A(b, p)))$$

Lastly, from around 18 months, ToMM appears. It is first manifested in pretend-play (Leslie, 1987). ToMM deploys metarepresentations:

Metarepresentation: $P_R(a, p)$

Here, I have used 'P_R' rather than the more common '*R*' to indicate that the relation that is represented is a fully intentional psychological relation. *a* is any agent, and *p* a proposition. Metarepresentations provide the child with the materials for representing her own psychological states as well as those of others. It takes some developing. Children first grasp such psychological states as pretend, know, and want, and then slowly come around to understanding such states as belief. This ability - the ability to metarepresent - is crucial for the acquisition of a folk psychological theory. For Baron-Cohen, as opposed to Leslie, ToMM doesn't just build metarepresentations, it also functions like a body of knowledge: (Baron-Cohen, 1995, pp. 54-55)

> "Children probably could also affirm a long list of axioms that constitute the core of their theory of mind, though as yet only a fraction of these have been explicitly stated and tested (such as "seeing leads to knowing," "appearance is not necessarily the same as reality," "people are attracted to things they want," and "people think that things are where they last saw them")."

This suggests the following view, which Baron-Cohen may or may not hold. The ID, EDD, and SAM don't just play a developmental role in the acquisition of folk psychological theory (which, we have just seen is located in the ToMM), they also make possible its proper application, once it is acquired. The ID provides an intuitive feel for what counts as action. The EDD directs you automatically to a source

It seems that only something like this is sufficient for *a* and *b* to share attention of *p*. Maybe the process goes on from here with increasing numbers of attendings (cf. Gómez, 1994), but I shall not go into this here. The point is simply that much more recursive attending is needed than what Baron-Cohen would lead us to believe.

Secondly, Baron-Cohen thinks that triadic representations are necessary for the construction of metarepresentations. Why this is so remains unclear. Why, e.g., doesn't the fact that we manipulate objects and we see other people manipulating those very same objects at different times give us the idea of a common point of view on the world in the absence of shared attention?

of information about psychological states, beliefs in particular. The SAM might give way to ToMM or remain in some form or other.[12]

As I have already indicated, and as Baron-Cohen's classification suggests, the ID, EDD, and SAM should not be regarded as *themselves* forming part of folk psychological theory. Certainly, the ID, EDD, and SAM all contain information, but they don't seem to form part of proto-folk psychological theory. They are mechanisms that allow us to latch onto aspects of reality that are psychologically relevant.

Before ending, I should not that there is good evidence that high performing individuals with autism or Asperger Syndrome learn at least part of the core folk psychological theory (Happé, 1994 & 1995), even though, according to Baron-Cohen, their SAM is severely impaired. It is noticeable, however, that they never become fully fluent folk psychologists. Hence, it appears that SAM is not necessary for acquisition of a folk psychological theory. In fact, it is not unlikely that neither ID, EDD, or SAM are necessary for the acquisition of such a theory, although they might greatly facilitate it.

f. Fodor who sses the pl perform distinction.

When can a child be attributed knowledge of folk psychological theory? As we have seen, the acquisition of folk psychological theory is a gradual process. Some theory theorists are happy to attribute neonates some form of folk psychological theory. However, folk psychological theory as we know it as adults, is obviously a much later development. It is acquired in different stages, and it seems fair to say that around the age of 4, children possess an important part of that theory - the idea that psychological states are representational states. And it seems that even before that, children had some idea about the possible causal connections between for example desire and action. Children don't understand lies and miscommunication until they are between 6 and 7, and a full understanding of intention follows that. However, if one understands the core of folk psychological theory as broadly as I do, it seems more safe to assume that children possess

[12] There is an obvious parallel here to Chomsky's work. According to him, we are born with knowledge of a Universal Grammar that develops into a particular grammar with experience. Such a process is known as parameter setting. Knowledge of UG allows us to develop grammars that put us in a position to understand grammatical sentences. This is the diachronic account of the function of UG. Once the parameters are set, this system does not become obsolete, but continues to function in its "grown" state in the production and comprehension of language (Chomsky, 1975). What UG has developed into, serves a distinctly synchronic function also.

knowledge of it when they are young teenagers. At the moment, this is as precise as we can be.

10. The Simulation Theory

As I am only concerned with the Simulation Theory insofar as it directly bears on the formulation of the Theory Theory thesis, the aim is not to present it in full detail, but to give a concise overview of the position. I shall leave out certain strands of simulationism, and I will not mention Diachronic Simulation Theory in any detail.

Being another internal account of folk psychology, the Simulation Theory must hold that simulation, rather than theory, is causally efficacious in the production of our folk psychological attributions. The idea of the first simulationists was that folk psychology is really based on knowledge-how, not on knowledge-that (Heal, 1986; Gordon, 1992b). Most simulationists have now abandoned the idea of simulation being necessary for psychological attributions. What is maintained is that it plays a large role in such attributions. There may even be part of psychological attributions where simulation plays an overarching role, for example attributions that involve inferring what someone thinks from what else they think (Heal, 1995).

Simulation is a widespread phenomenon. Computers are used for simulations of anything from the behaviour of manmade objects (in engineering, for example) to human reasoning (AI studies). In aeronautics, wind tunnels have been used to test the flight patterns of aeroplanes. Wind tunnels are small scale atmospheres where miniature planes are exposed to various atmospheric phenomena. Testing miniature planes in wind tunnels is a simpler way of gaining information about the capacities of an aeroplane than calculating it using available theories. The only calculation required in wind tunnel testing is that of scaling up from the miniature environment. A case of simulation such as this provides a model for mental simulation.

In mental simulation minds simulate other minds. It is different from our wind tunnel example in that the system that carries out the simulation is also the one that supervises it, reads off the result, and draws the relevant conclusion. In aeronautics this role is played by an engineer. Furthermore, the system that carries out the simulation is a system of the same kind as that which it is simulating. It is not simply that a mind is simulating a mind. Simulation deploys some form of reasoning procedure in one system to determine what reasoning is carried out in another system (but see Rational Simulationism below). Mental processes are deployed to imitate other mental processes of the same type - for example hypothetical reasoning, decision making, and belief formation. So, mental simulation has certain advantages over other kinds of simulation since minds have a great number of things in common. Simulationists assume that they have enough in common to make mental simulation a relatively precise and useful tool for understanding ourselves and others. This so-called 'assumption of similarity' applies to most aspects of mental functioning, such as theoretical and practical reasoning, and the formation of beliefs, desires, and emotions.

One of the most commonly used examples of simulation, is that of a decision making process. We simulate this by doing whatever it is that we do when we ourselves make decisions. We, ourselves, are the model we use in simulating. Robert Gordon calls this the Model Model of simulation (1992a, p. 117). Some philosophers believe that we have a decision making system that we deploy in such simulations (see Goldman, 1995; Stich & Nichols, 1992 and 1995). When we use it to make decisions, it is said that it is used on-line, and when we use it to simulate, we use it off-line. This is why simulation is sometimes known as off-line simulation, as in the title of Shaun Nichols, Steven Stich, Alan Leslie & David Klein's paper: "Varieties of Off-Line Simulation" (1996).

'On-line' and 'off-line' were originally computer terms. A computer that operates on-line, operates "under the direct control of, or connected to, a main computer" (Random House Webster's Unabridged Dictionary). A computer running off-line is not connected in this fashion to a main computer. So, the idea behind using these terms in Simulation Theory, is to indicate that when we simulate, the result of the decision making process does not have the effects it usually does - it does not dispose the agent to make a decision, form an intention, or act in a particular way. Rather, it furnishes her with information. This information, in its turn, may well have important effects on the agent's behaviour. However, the effects of using one's decision making system on-line are very different from those of using it off-line. Deciding whether to go to Spain or Germany over the summer will generally lead me to decide to do one or the other. That, in its turn, will significantly increase the probability of me going to the country I've decided to go to. *Pretending* to decide whether to go to Spain or Germany over the summer will have no such effects.

Another way of looking at simulation is more in terms of hypothetical reasoning (Gordon, 1992b; Heal, 1994b, 1995 & 1998; Davies & Stone, 1998). This is not to say that decision making no longer plays any role, but that where decision making does play a role in simulation, it is understood broadly in terms of decision making procedures, rather in terms of a system or systems. It is a view of simulation that doesn't give the impression that off-line simulation does, namely that simulation is quite automatic and effortless. No doubt, it sometimes is. But by stressing hypothetical reasoning, these simulationists stress that simulating might take some effort and need not be regarded as a kind of automatic process. Nevertheless, the profile in both Simulation Theories is markedly different from Theory Theory. Here is no inference based on folk psychological generalisations. What is relied on is a capacity for figuring out what one would do under the kind of circumstances that the agent is in, in the context of a simulation. The Off-Line Simulation Theory stresses process, Hypothetical Reasoning Simulation Theory stresses reasoning procedures and/or rules for reasoning. I will go into more detail with the rules of reasoning approach in chapter 2. Most simulationists, however, are willing to agree that there is *some* kind of knowledge base concerning psychology that is drawn upon in simulation. It is just that this knowledge base is significantly different from folk psychological theory.

A central idea for many simulationists is what is known as the 'assumption of similarity' (Goldman, 1989; Heal, 1986). To explain this, we need to look a bit closer at how simulation works. Imagine you want to predict what someone is going to do. Imagine that the situation is ideal and you are in possession of knowledge concerning their relevant beliefs and desires. Take John again; you know that he wants a great garden party and he fears that it will rain. In order to simulate John, you simply imagine that *you* have these beliefs and desires and work out what *you* would do. For example, you imaginatively decide to move the tables and chairs inside, if possible. Once you have imaginatively decided what to do, you infer that the person whom you are predicting would do just that. What underlies this inference is the assumption of similarity. You assume that given the same beliefs and the same desires, any agent would make the same decision. Jane Heal formulates the idea thus: (Heal, 1986)

> Only one simple assumption is needed: that they are like me in being thinkers, that they possess the same fundamental cognitive capacities and propensities that I do.[13]

It is perhaps clearer how this assumption works if we imagine simulating what an agent will think on the basis of what they think now. Imagine that they believe that if *p* then *q*, and they also believe that *p*; for example, if it rains, then the garden party will be ruined, and it is going to rain. We pretend that we have just those beliefs, and ask ourselves what else we would believe in that situation, coming up with the pretend-belief that *q*; the garden party will be ruined. Certainly here, the assumption of similarity seems eminently reasonable.

The assumption of similarity is crucial on this picture. If we know that the agents that we are interested in understanding have certain psychopathologies, for example, I will have to adjust the assumption. Normally, even seriously disturbed people are comprehensible to a degree. Someone who believes that there are little green men coming out of the electric sockets is incomprehensible *vis-à-vis* that belief. However, we can explain their putting sticking plaster over the sockets, for example, by reference to their belief. Here, we cannot use simulation to *get* at their belief because the assumption of similarity has broken down, but we can to some degree simulate what they will do given knowledge of their beliefs. However, if I am really interested in understanding mentally disturbed people, it may be wise for me to resort to the best psychological theory. It is an interesting

[13] There are certain dangers to presenting the Simulation Theory as if it were a unified position. In fact, just like Theory Theory, the Simulation Theory covers a group of positions with certain family resemblances. Heal is at the more rationalistic end of the spectrum. Talking about 'cognitive capacities and propensities' seems to indicate that the assumption of similarity does not cover agent's emotional and affective lives. For many simulationists, such as Goldman and Gordon, it does.

question whether disturbed people are better at understanding other people that are disturbed in the same way.

The assumption of similarity also concerns the similarity between pretend psychological states and *bona fide* psychological states. For simulationism to work, it must be the case that pretend psychological states have similar causal powers to *bona fide* psychological states. In fact, if the tokening of the belief that *p* causes the tokening of the belief that *q*, given the belief that if *p* then *q*, then the tokening of the pretend-belief that *p* must cause the tokening of the pretend-belief that *q*, given the pretend-belief that if *p* then *q*. The basic difference between the two, is that the causal chain is constituted by psychological states in the one case and by pretend-psychological states in the other. Of course, pretend-psychological states are themselves psychological states, but they are psychological states of a different type from the *bona fide* psychological states that they are imitating. The same goes for a simulated decision making. The pretend-desire that *q* and pretend-belief that if *p* then *q*, ought to give rise to a pretend-decision (say, the pretend-decision to attempt to bring it about that *p*) that corresponds to the decision that the desire that *q* and the belief that if *p* then *q* would give rise to (say, the decision to attempt to bring it about that *p*).

Or the belief that if p then q

Some psychological states seem more likely candidates for simulation than others - beliefs, for example. We often engage in counterfactual reasoning and this often proves useful. For example, in determining which of a number of different tools to use to reach the apples on my apple tree, instead of trying each on out in reality, I can try them out in imagination. I pretend-believe that I'm holding the rake, and I pretend-see whether it is long enough to reach the apples I want to pluck. I, as it were, pretend-see that it isn't, on which basis I conclude that the rake, by itself, won't get me the coveted apples. Assuming that my ideas of the distance to the apples and the length of the rake are largely correct, pretend-believing that I am trying to reach the apples has similar causal powers to actually believing that I am trying to reach the apples.

What is pretend-seeing?

On the other hand, it seems less obvious that imagining being in affective states has similar causal powers to actually being in these affective states. It may be difficult to get oneself into an affective state, or perhaps some affective states, such as grief and rage, are so

empathy?, fiction?

Donadio

unpleasant that one is not prone to get into such states simply to gain understanding of others. Alternatively, it may turn out that unless one imagines (some?) affective states very vividly, the corresponding pretend-states won't have the requisite causal powers. And if this is the case, one might argue that if one's imaginings have to be so vivid that one actually experiences the affect, it is no longer a simulation, it is no longer a pretend-affective state. All this deserves closer attention. My point is not that simulation will only work for beliefs, I don't think that I am in a position to say that. It is simply that whereas belief is a good clear case for simulation, other psychological states seem less so.

The assumption of similarity also plays a role in gathering information about people's psychological states in the absence of any prior or relevant information. We need material with which to start the simulation - just as we need material on which to apply our folk psychological theory. In a predictive simulation, we can derive it from what the subject has said or done or what her environment is like. Here, we imagine being in the relevant environment and having done or said what the subject has done or said.[14] We should then get to have certain pretend-psychological states that can be used to simulate the subject's future thoughts or actions.

An important function of the assumption of similarity is that of justifying all the different uses simulation is put to, for example, the use of pretend-psychological states achieved by imaginative identification in a simulation. These psychological states form the basis of a prediction of a subject's actions or thoughts, and only if it is reasonable to suppose that the subject did possess such states will it be reasonable to make the relevant prediction. The assumption of similarity may also function as a motivation for simulating subjects. After all, if subjects had no reason to think that their pretend-states mirror real states of others, simulation would lose its significance, and hence would be reduced a sport, at best, for all but instrumentalists.

The assumption of similarity differs from Theory Theory's generalised statements concerning human psychology in the following way. Theory Theory is always completely explicit about the similarities

[14] Alvin Goldman (1989) has suggested that simulation can account for the interpretation of language. He opposes his simulationist view of interpretation with, among others, Donald Davidson's radical interpretation (Davidson, 1984). What difficulties this might provide for the simulator that will have even less material with which to start a simulation, I cannot explore here.

across persons. Thus, if a person believes that if it is raining, the streets are wet, and also believes that it is raining, then that person will believe, or come to believe, that the streets are wet. Folk psychological theory contains statements that generalise either over all (rational) human beings (cf. chapter 2, section 5) or specify the relevant characteristics of the group at issue, when it comes to character, for example. So, people are alike in exactly the ways specified by these generalisations.

Depending on the particular version of the Simulation Theory, how the similarity between people is understood is more vague. Alvin Goldman suggests that we spell out this similarity in terms of "psychological preferences for certain modes of categorisation and 'entification'" (1995a, p. 90). Heal talks at various points of "cognitive competence" (1986) and "rationality" (1996). Whereas Goldman's idea seems to be almost unnecessarily narrow, Heal's is, perhaps, too broad, as Stich & Nichols (1997) complain. It is important, though, to point out that all simulationists allow for circumstances where the assumption of similarity doesn't hold, and most agree that knowledge of these circumstances is theoretical - know-that. Heal talks of the assumption as a "projectivist first move" (Heal, 1995, p. 49), which needs to be revised in the light of information about a number of factors, such as visual perspective and educational background. This information is theoretical in nature since knowledge of the constraints of the respects in which we, as agents, are alike in forming psychological states amounts to theoretical knowledge of psychological factors on at least one understanding of 'theoretical knowledge'. Therefore, we should expect theoretical knowledge about psychology to play *some* role in folk psychological attribution.[15] However, this doesn't make the knowledge in question like that of folk psychological theory because not only is it much less encompassing, but it must also always be combined with some simulating activity to produce psychological attributions.

Some simulationists resist the idea that simulation relies on an assumption of similarity. Gordon believes that we can do away with it

[15] It is the fact that we must also rely on theoretical knowledge of (folk) psychology that is at issue here. A simulation can draw on any theoretical knowledge provided that it is not psychological, consonant with it being a distinct position from the Theory Theory.

altogether. It is unnecessary, he claims, since when we simulate we do not pretend that we are someone else and then attribute to them whatever states result from such a pretence. Rather, imagining being someone else involves an ego-centric shift whereby the result of the simulation automatically applies to the subject with whom we are identified. For example, in the decision: "I am going to complain to the highest authority", the 'I' refers not to ourselves but to the simulated subject (1992a and 1995).[16] 1998?

Martin Davies and Tony Stone (1988) have suggested that the respect in which subjects are similar is in respect of reasoning correctly. The assumption of similarity becomes an assumption of *right reasoning*. That is, in a simulation we deploy a normative idea about what reasoning is correct (as opposed to just how I happen to reason) and assume that all subjects reason correctly (*ceteris paribus*, naturally). This is a variation on the simulation theme (but see chapter 2, section 6-7). However, as the authors acknowledge, there are certain limits to such an approach since it can only be used to explain, predict and understand what the right thing to do, think, feel, and act is. And there are certainly cases, where it is not so clear that this can be done. Is there a *right* thing to feel under certain circumstances? Is feeling relief when one has narrowly escaped a dangerous situation a case that can be simulated in this manner? So, this account has certain limits built into it.

11. Simulationist Accounts of Folk Psychological Explanation

Simulationists have traditionally concentrated their exposition of the Simulation Theory on psychological prediction and identification of psychological states. Some simulationists, like Heal, think that there is only a limited role for simulation is psychological explanation. Others, like Gordon, see no problem with extending simulation to explanation. Heal has suggested that the main role of simulation is played in prediction of thought and action based on prior knowledge of thoughts (Heal, 1995). On the other hand, she clearly believes that simulation *can* be effective in producing psychological explanations (Heal, 1998).

[16] A problem for Gordon's account is that whereas an assumption of similarity provides justification and motivation for simulating, an ego-centric shift does neither.

What I must do is to work back from a particular behaviour to the psychological states that caused it, along the lines of: "She pulled a funny face: was she really amused?" (1998, p. 86). This appears similar to Gordon's account of explanation. He thinks that we look at the subject's environment to discover salient features that may have influenced the agent. For example, if the agent believes that she is being followed by government agents, we look to see what in her environment might have given her that idea. So we put ourselves in the environment of the agent prior to them thinking, wanting, doing, and so on, what we want to explain. From here, we proceed as in cases of prediction (1992a). That is, we imagine the kinds of thoughts, wants, and so on, we would have under such circumstances, and on this basis decide what we would do, think, and so on. If we pretend-decided to act just as the subject did, then we attribute the preceding pretend-psychological states to the agent. Whereas prediction-simulation uses our decision making capacities, explanation-simulation cannot do so because decision making is always a forward process - it concerns what we are going to do, think, etc., in the, sometimes very near, future. Explanation works backwards. Hence, it must do with some more general notion of hypothetical reasoning. But hypothetical reasoning isn't as nicely tailored for explanation by simulation as decision making is for prediction by simulation.

not quite

?

A hurdle simulationists have met is the idea that explanation is deductive-nomological in nature. According to this view, something is only an explanation if it contains in it a reference to some law or generalisation that subsumes the explanandum (Hempel, 1965). It is often assumed that Theory Theory is committed to such an account of explanation (but see chapter 3, section 1). However, if this view is accepted, then it follows directly that only one strand of Simulation Theory can provide us with folk psychological explanations: Rational Simulationism. For this version allows us to refer back to normative rules under which a particular thought-process or behaviour would fall. Simulationists have therefore been concerned to develop an alternative account of explanation.

Aim is to model the sorts of expl we, we folk, tend to give. And they do not include covering laws.

Gordon has suggested that if one models the particular situation that one needs to explain, one can explain - in a reasonable sense of that word - what happened by reference to that modelling. The explanation explains by picking out the relevant cause(s). If one is also interested in knowing *why* these causes were productive of the situation at hand, one might have to point back to some law. But not in the case of psychological explanation. By running through, in imagination, a similar process to that which the agent went through, simulation allows us to see/understand from the inside, how the agent got to act (or think, feel, etc.) the way she did. We come to see "*the relative attractiveness we generally see in our own actions at the time we act.* The explanatory understanding that had eluded you before is thus *empathic* understanding." (Gordon, 1992a, p. 117).

Other simulationists are even more interested than Gordon in highlighting the difference between scientific explanations and psychological ones. Heal claims that: (1986, p. 52)

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

For Davies & Stone, also, a simulation-explanation provides a *first-personal* understanding of the subject (1998). In some sense, we come to know what it was like for the subject. Heal also talks of how we manage to capture the world from a particular point of view (1998). It seems relatively clear, that due to its empathic, first-personal nature, simulation provides a good strong case for psychological *understanding*. Whether the explanatoriness of psychological explanation can be explained thus, is a further question.

12. Simulation and Self-Ascription of Psychological States

Lastly, let us consider how simulationists regard self-attribution of psychological states. On the face of it, it appears in order to be able to simulate at all, we must first be able to attribute psychological states to ourselves. We must be able to self-attribute the psychological states that we imagine the agent possessing, we must be able to reflect on them, be they pretend or *bona fide* states, we must be able to know what state our simulation leads us to be in, and what decision we have made, for example. In other words, the ability to attribute psychological states to ourselves is *prior* to being able to attribute such states to anybody else in the following sense: if we could not self-attribute psychological states, we wouldn't be able to simulate at all.

Alvin Goldman agrees that self-attribution must precede simulation (Goldman, 1989). He is happy splitting folk psychological attribution into two: attributions to self and attributions to others. The latter is asymmetrically dependent on the former, and the Simulation Theory only accounts for the latter. This is not a special problem for simulationism, he claims, for Theory Theory provides no satisfactory account of first person psychological attribution either (but see chapter 5). Other simulationists are less insouciant about such a split. Heal maintains that simulation is involved in the attribution of first personal psychological states also (1986). This approach is somewhat mysterious since if we simulate ourselves, it seems that we must be able to identify the result of the simulation before we can attribute this state to ourselves. But this, of course, is impossible.

Gordon views the issue of self- and other-attribution of psychological states as a kind of boot-strapping process. You need a little bit of one to get the other, and then you move on from there until you've got full-blown simulation, where there is no real saying which attributions rely on which. He gives a detailed ontogenetic account of the ability to simulate (Gordon, 1995). We teach children to preface their requests, for example 'chocolate', with 'I want' such that they learn to refer to their desires - 'I want chocolate' - even before they have a concept of desire. In this fashion, they learn to self-ascribe psychological states whilst having no concept of them. So, training external to simulation provides them with reliable non-comprehending self-ascriptions, on the plausible assumption that there is more to possessing a concept than being able to apply it reliably under the right circumstances. This allows them to begin some kind of simulation - say proto-simulation. Through experience with this, and in the course of development of other abilities, the child will come to master psychological concepts. She will be able to apply these to the states she already knows how to pick out. Once this has occurred, she will be able to engage in full-blown simulation.

With respect to belief, children first learn to identify this using an Evansian ascent routine (Evans, 1982). When posed with a question as to what beliefs she holds, say: "Do you think that the cookie is in the cupboard?" the child simply rephrases the question as "Is the cookie in the cupboard?" and answers this question. She need have no concept of belief in order to answer questions about her beliefs. As with the case of desire, this training will allow her to begin to simulate and with practice she will gain a concept of belief that she will then attach to her ascent routine. In both the cases of desire and belief, development takes place by first encouraging the child to identify some internal state that allows for proto-simulation, a concept is then acquired through such simulation, after which the child can properly simulate.

In conclusion, with respect of self-attribution of psychological states, there are a number of simulationist options. How Theory Theory accommodates self-attribution is the topic of chapter 4, and we will see that here, also, there is room for some variation.

13. Conclusion

We have now been introduced to folk psychology and the two prevalent internal theories of it. In some respects, the description of the Theory Theory has been more elaborate than that of the Simulation Theory - giving many examples of generalisations, *ceteris paribus* clauses, and how they work in reasoning - but in others it has been less so. What about self-knowledge, for example? In the next four chapters, I shall explore the following aspects of Theory Theory: what does it mean to say that we have knowledge of a folk psychological *theory* (chapter 2 & 3)? Is Theory Theory a functionalist theory (chapter 4)? Can Theory Theory account for self-knowledge (chapter 4)? And is our knowledge of folk psychological theory tacit (chapter 5)?

Chapter 2

Folk Psychology & Folk Theory

> But, of course, there are lots of domains of commonsense knowledge in which it is rather implausible to suppose that the mentally represented "knowledge structure" includes theoretical constructs linked together in law-like ways. Knowledge of cooking or of current affairs are likely candidates here, as is the knowledge that underlies our judgments about what is polite and impolite in our culture. And it is entirely likely that folk psychological knowledge will turn out to resemble the knowledge structures underlying cooking or politeness judgments rather than the knowledge structures that underlie the scientific predictions and explanations produced by a competent physicist or chemist. [...] On the inclusive reading of 'theory', any mentally represented body of information about a domain counts as a theory, regardless of how the information is encoded or whether it includes theoretical constructs or nomological generalizations. (Stich & Nichols, 1996, pp. 146-7)

The knowledge that we are all[17] said to possess in the absence of training or specialisation, is sometimes called everyday, common sense, or folk knowledge. Other than folk psychological knowledge, folk biological and folk physical knowledge have received most attention. Recent years have seen a flourishing of research on everyday knowledge in philosophy, ? physics? psychology, anthropology, and sociology (Atran, 1994; Carey, 1985; diSessa, 1988; Keil, 1994; McCloskey, 1983; Semin & Gergen, 1990). In all of these areas of everyday knowledge, there is disagreement about how best to characterise our ability to explain, predict, and understand the relevant phenomena. Is it best described as knowledge of a theory - and an ability to use it - or as something else? Simulation is most plausible only as an alternative to theory in folk psychological knowledge. Embodying a mind puts on in a good position to simulate other minds, but not to simulate physical or biological phenomena. In folk physics and biology, the discussion mainly concerns whether the requisite body of knowledge that we draw on is a theory or some more disparate collection of principles and rules of thumb (but see Harris, 1994).

Folk knowledge provides a good starting point for our exploration into how best to characterise the theoreticity of folk psychological knowledge. We can here see the different views of theoreticity at play in a very general way. There are at least two very common usages of the term 'theory'. One is pretty loose. Principles concerning some subject matter, for example cooking, might count as a theory on this account. We mostly have this usage in mind when we say things like: "I've got a theory of how that works". The other usage is stricter. It is used to capture some more well-defined and systematic body of laws or principles. We have this usage in mind when we talk of quantum theory, personality theory, Bayesean theory, and so on.

The discussion in folk knowledge concerning whether this knowledge is knowledge of a theory, is characterised by a disagreement

[17] The knowledge that all normal subjects are said to possess. For people with autism or Asperger Syndrome may not possess folk psychological knowledge - and if they do, the acquisition of it takes considerably longer than for normal subjects.

schiz?

about which of the two common senses of 'theory' is to be deployed. It is therefore crucial for theory theorists to decide which meaning to adopt such as to properly shape future research and debate. In the next two chapters, I shall examine these two possibilities. Here we will be concerned with the loose sense - I shall call it the Folk Theory Theory, and in chapter 3, I will move on to the stricter sense.

What I will do here is the following. First, I will present just how the Theory Theory debate runs in folk physics and folk biology. I will then present a suggestion that we should understand 'theory' loosely. The problem that immediately arises is that such a loose understanding might lead to the collapse of the Theory Theory versus Simulation Theory debate. However, there is a redescription of the debate in terms of mental representations, that will serve to save the debate from collapse. I call this the minimal distinction. In the process of arguing for this distinction, one version of Simulation Theory that doesn't seem to lend itself to such redescription is examined. On closer inspection, it turns out not to be a *bona fide* Simulation Theory, hence does not constitute a threat to the distinction. Despite the fact that Folk Theory Theory won't lead to a collapse of the debate, I will not champion this position. I conclude that it too vague and uninformative to helpfully shape the debate within folk psychology.

1. Folk Theories

Michael McCloskey believes that various experiments support the idea that a version of the medieval impetus theory of motion forms part of our folk physical knowledge: (McCloskey, 1983, p. 306)

> First, the theory asserts that the act of setting an object in motion imparts to the object an internal force or "impetus" that serves to maintain the motion. Second, the theory assumes that a moving object's impetus gradually dissipates (either spontaneously or as a result of external influences), and as a consequence the object gradually slows down and comes to a stop.

Presumably other small theories, for example that of centrifugal force, also form part of this body of knowledge. At any rate, knowledge of this naive theory of motion is causally efficacious in the production of our everyday judgements about the motion of physical objects. As I understand McCloskey, knowledge of the impetus theory is *necessary* for the production of our folk physical judgements, or at least a subclass of these judgements. However, I don't think he believes that it is sufficient. Looked at in this way, folk physics forms a complement to folk psychology.

In McCloskey's experiments, the impetus theory was only one piece of knowledge among others, that was productive of subjects' judgements. In addition to naive theories, subjects make use of: (1983, p. 321)

> analogies, memories for specific experiences (e.g., throwing a rock with a sling), isolated facts about mechanics (e.g., Galileo found that heavy and light objects fall at the same rate) and knowledge acquired through formal instruction in physics (e.g., a projectile's motion can be analyzed into independent horizontal and vertical components).

I think this is a general feature of folk knowledge. Not necessarily that it is composite, but that it is often used alongside other information. Think of folk psychological knowledge. Folk psychology is a heterogeneous domain where we use not only folk psychological theory, but whatever other knowledge comes in handy; for example knowledge of folk physics, social mores, experimental psychology, and psycho-analysis, as already mentioned in chapter 1. A difference is be that whereas knowledge of experimental psychology and psycho-analysis can be incorporated into folk psychological theory, knowledge of folk physics and social mores are more likely to stay external to it.

Another way of regarding naive physics can be found in Andrea diSessa's work. He has the following to say about McCloskey-type folk physics: (diSessa, 1988, p. 50)

> this is a highly misleading representation of the actual state of affairs. Though it gives signs of being quite robust, intuitive physics is nothing much like a theory in the way one uses that word to describe theories in the history of science or professional practice. Instead, intuitive physics is a fragmented collection of ideas, loosely connected

> and reinforcing, having none of the commitment or systematicity that one attributes to theories.

According to diSessa, the evidence does not support the claim that we have knowledge of a naive impetus theory. Subjects don't always seem to use this theory in their predictions or explanations of objects moving; at least not in any obvious or simple way. This, however, is what we should expect if McCloskey is right that knowledge of naive impetus theory is necessary but not sufficient for such predictions and explanations. diSessa also thinks that many other mini-theories must be added to the impetus theory in order to explain people's judgements about motion. However, if this is a point about sufficiency, it is not a criticism of McCloskey. In any case, diSessa is not optimistic about the prospects for a folk *theory* of motion. Rather, he suggests that our naive knowledge really just is "knowledge in pieces". These pieces are not integrated with each other and are not deep and explanatory but are simple abstractions from everyday experience.

The situation looks much the same in folk biology. Carey, for example, has argued that we have knowledge of a theory of biological categories (Carey, 1985). What we know is a theory because it has certain important features in common with scientific theories. It is "characterised by the phenomena in its domain, its laws and other explanatory mechanisms, and the concepts that articulate the laws and the representations of the phenomena" (p. 201). As opposed to this, Scott Atran has argued that our folk biology is not properly regarded as a theory. This is due to the fact that different cultures have different explicit ideas of aspects of biological function, say reproduction, and yet taxonomize in very similar ways to each other. This is best explained, Atran claims, by assuming that "the categorical structure of living kinds, including plants and animals, [are] the product of domain-specific processes that are largely theory- and culture-independent." (Atran, 1994, p. 334).

Two important assumptions lie behind Atran's argument. One is that folk biology is culture-universal - all cultures taxonomize in similar ways. Therefore, if the ability to taxonomize is derived from knowledge of folk biological categories, then all cultures must possess the same knowledge. But if there is a folk biological theory that we all have knowledge of, our explicit ideas about biological function should be importantly influenced by this knowledge. In other words, we should have some kind of coherent theory of biology where explicit and implicit ideas cohere (our knowledge of folk biology being implicit). However, that is not the case. Or, so the argument goes.

The second assumption concerns our implicit knowledge of folk biology. But by arguing the way that he does, Atran rules out the possibility that implicit knowledge of folk biology could be similar to knowledge of grammar (cf. chapter 1). According to Chomsky, our explicit ideas of grammar sometimes clash with our tacit knowledge of grammar. This might be exactly what someone like Carey has in mind. Our taxonomizing ability might derive from our tacit knowledge of folk biological theory, our explicit theory being different. Another way to look at it is to say that Carey could be providing an internal account of biological classification, and what each culture explicitly holds are external accounts of the same subject matter. Atran seems to assume that we cannot regard our knowledge of taxonomy as knowledge of a theory, but I don't believe that he has given us any reason to discount this option.

In this context, it is worth stressing that the idea that the core of folk psychological theory is culture universal, has some following; what is acquired seems to be the same, and it seems to be acquired in the same order (Astington, 1994). I imagine that the corresponding idea has following among simulationists also. However, it is outside the scope of this thesis to try to determine the truth of this idea.

Some evidence against cultural uniformity of FP & its acquisition

Looking at the competing views in both folk physics and folk biology, one thing becomes immediately clear: different notions of 'theory' are at play here. Whereas McCloskey seems to have something pretty broad in mind when he talks of 'theory', Atran, Carey and diSessa's think of 'theory' in a much narrower way. So, whilst Atran and Carey disagree about whether folk biology is a theory in the same sense of 'theory', McCloskey and diSessa don't seem to do so. So, it appears that to tidy up the debate in all areas of folk knowledge, we need to agree on one use of 'theory'. We might, for example, decide to use 'theory' broadly.

2. Folk Theory Theory

In the introductory quotation to this chapter, we have Stich & Nichols presenting their idea of what they call the inclusive use of the term 'theory', and I have called the 'broad use'. Since "any mentally represented body of information about a domain counts as a theory" (1996, p. 147), knowledge of folk physics, folk biology, and folk psychology can all be regarded as knowledge of theories. We might call such theories 'folk theories'. According to Stich & Nichols, Theory Theory is best understood as claiming that we have knowledge of a folk theory, rather than something like a scientific theory.

I think they intended to be uncommitted on this.

This way of regarding a theory seems to make it more plausible that our folk knowledge is knowledge of theories. There certainly seem to be sufficient dissimilarities between scientific knowledge and folk knowledge to make it implausible that both are knowledge of theories in the same sense. We don't experiment rigorously, we don't take extreme care that our conceptual framework is coherent, and so on. Also, we are not being pedantic about word-use.

inclusive sense?

There are, however, certain problems with this very inclusive reading of 'theory'. One is that it has become too inclusive. The term has lost its sharpness, and using it to describe a body of knowledge will not give much of an insight into the structure of this body. By saying that something is a theory, all I am saying is that it is a body of information that is related in some way to a domain - presumably the component concepts are interrelated in some interesting way. Instead of having a relatively tight notion of theory that imparts a lot of information, we end up with a loose and impoverished notion.

so scientific theories are folk theories?

Another problem that is closely connected to the first one, is that this way of conceiving 'theory' is in danger of blurring the distinction between Theory Theory and Simulation Theory. Simulation theorists have always been somewhat sensitive to how theory theorists define their position. And rightly so. Claiming that our knowledge of folk psychology is knowledge of a folk theory is quite different from claiming that it is knowledge of a proto-scientific theory. If Theory Theory becomes Folk Theory Theory, there is very little room for simulationism, and certainly none for anyone who, like diSessa, wants to characterise folk psychological knowledge as being importantly different from theoretical knowledge.

explain.

For the time being, let us leave the first objection to the side, and see if we can actually make Folk Theory Theory fly. To do that requires that it leaves enough room for the Simulation Theory, such that Theory Theory doesn't become vacuously true. That it does so, is not immediately obvious. In fact, it seems not to.

3. The Threat of Collapse

It is a significant problem that on a particular way of construing the Simulation Theory, it is at risk from collapsing into a particular construal of the Folk Theory Theory. Davies and Heal have both voiced concern about theory theorists claiming that we have *tacit* knowledge of folk psychological theory (Davies, 1994; Heal, 1994b). The problem is the following. A good theory mirrors what it is a theory of closely: (Heal, 1994b, p. 131)

> a good *explicit* theory enables us to produce an unfolding sequence of representations which runs parallel to developments in the item to be understood.

A good theory of how a heart functions, for example, will closely mirror the functioning of a heart.[18] The derivational structure of the theory of the heart will closely mirror the causal structure of the heart. If I want to predict how a heart will react to a particular pattern of stimulation, say, I can either deploy the theory, or I can stimulate another heart in the relevant fashion and see how it reacts. Granted, there might not be much point to this, but it is certainly a possibility open to me. Now, imagine the same situation with respect of folk psychology.

If folk psychological theory is a good theory of how people think and act, its derivational structure will closely mirror the causal structure of how people think and act. To see what is meant here by derivational structure, think back to the use of the action generalisation in chapter 1. There we saw how we can derive what a person will attempt to do on the basis of knowledge of (G1) and of what they want and what they believe, provided that *ceteris* are *paribus*.

[18] Whereas this may be true of theories of hearts and hearts, it is more difficult to see why it is true of scientific theories in general. How, for example, will a derivation from Newton's Laws of Motion mirror the movement of the body in question?

Now, presumably the causal pattern in the person is something like this: they want *q*, they believe that if *p* then *q*, and, *ceteris* being *paribus*, (G1) is a law that governs their decision making. So, they will attempt to bring it about that *p*. Here it is relatively clear that the derivational structure of folk psychological theory mirrors the causal structure of the reasoning and the decision making systems (or whatever subserves such functioning).

One way of explicating the notion of tacit knowledge is in terms of the match between derivational and causal structure: (Davies, 1994, p. 115)

> Roughly speaking, a component processing mechanism embodies tacit knowledge of a particular rule or axiom if it plays a role in mediating causally between representational states that is structurally analogous to the role that the rule or axiom itself plays in mediating derivationally between premises and conclusions...

If we accept this view of tacit knowledge, then if the Theory Theory holds that we have tacit knowledge of folk psychological theory, then the collapse of the Theory Theory versus Simulation Theory debate would appear to ensue. It may turn out to be impossible to distinguish between a predictor using folk psychological theory and a predictor simulating. A simulator harnesses her own reasoning and/or decision making system in deciding what the relevant agent will do. A folk psychological theorist uses her theory of reasoning or/and decision making to do so. However, if this theorist's knowledge of the theory amounts to the causal structure of her psychological processes matching the derivational structure of the relevant theory (because the theory is tacitly known), we are in trouble. As we saw above, this causal structure may be indistinguishable from the causal structure that can be observed in the simulationist, since if the theory is a good one, its derivational structure ought to match the causal structure of the relevant state of affairs. And the simulationist is going through the relevant state of affairs albeit in imagination. In short, at the level of causal processes, a person using a folk psychological theory in her predictions (explanations/understandings) may be indistinguishable from those of a simulationist simulating her way to her predictions (explanations/understandings). In this case the debate collapses.

Davies (1994) has suggested that the threat of collapse relies on a particular, and wrong headed, picture of the Simulation Theory. The crucial point is this. When we simulate, the contents of our mental states are not *I believe that p, I desire that q,* and so on. Rather, they are: *p, q*. There is no need for psychological states to be represented in this process, for psychological states in the pretend mode are themselves causally efficacious in the simulation. This differs from the Theory Theory position, where the psychological states that are involved in reasoning about others must have the contents: *a believes that p, a desire that q*. That is, psychological states must be explicitly represented in the reasoning process in the Theory Theory case. To see how this might save the debate from collapse, we need to develop the proposal a bit.

4. The Minimal Distinction

I dub what is supposed to save the debate from collapse 'the minimal distinction' because it restricts both theories as little as possible whilst maintaining a difference between them. The distinction is set at the level of cognitive processes, and revolves around mental representations. It concerns the representational complexity involved in either theory. I will put the distinction in terms of functional architecture or, as it is also known, boxology. However, the distinction does not depend on accepting this way of regarding mental architecture. The crucial point is that one accepts that psychological states are representations of varying complexity.

We have already discussed mental representations in chapter 1. Boxology is a different way of thinking about the same thing. Here we talk of a representation being tokened in one of our psychological mode boxes, the belief box, the desire box, and so on. For example:

BELIEF BOX	INTENTION BOX	DESIRE BOX
the sky is grey	to read *Middlemarch*	I win the Nobel prize

This is a boxological way of representing the belief that the sky is grey, the intention to read *Middlemarch*, and the desire to win the Nobel prize. Awareness of having representations such as the above, is represented thus:

BELIEF BOX

I believe that the sky is grey
I intend to read *Middlemarch*
I desire to win the Nobel Prize

So, I believe that I believe that the sky is grey, that I intend to read *Middlemarch*, and that I desire to win the Nobel prize.

I shall call a mental representation of a mental representation, a 'mental metarepresentation'. A metarepresentation may represent our own representations or those of others. For example:

BELIEF BOX

Carol wants it to snow
Asger enjoys raping and pillaging
John fears that it will rain

Only through metarepresentations do we become aware of thoughts *as such*, be they our own or those of others. This much should be agreed on both sides.

Here's how we can recast the Theory Theory versus Simulation Theory debate in terms of mental representations. Theory Theory maintains that all folk psychological reasoning takes place in terms of metarepresentations. This is because psychological states form part of the contents of the thoughts that we have when we reason folk psychologically; for example 'a desires that q and believes that if p then q'. We have beliefs about agents and their psychological states. Simulationists, on the other hand, need not hold the same. They can make do with psychological reasoning taking place in terms of simple, or object-level, representations. This is because instead of representing

Heal points out that the simulator must begin with $B(FBp)$.

to themselves the psychological modes of the subjects that are simulated, simulators can themselves instantiate something close to these modes. They *have* a pretend belief or a pretend desire, they don't have beliefs *about* pretending to have a belief or a desire.

Take the following example of a piece of folk psychological reasoning:

> Abel wants to make Pip a gentleman, and he believes that if he makes sure that Pip has access to lots of money, he will become one; therefore Abel does his best to make sure that Pip has access to lots of money.

One way of looking at this process is that the following beliefs are consecutively tokened in the belief box as a consequences of engaging some sort of reasoning mechanism:

REASONING PROCESS

BELIEF BOX	BELIEF BOX	BELIEF BOX
(Input) Abel$_a$ wants to make Pip$_p$ a gentleman$_g$	(Input) Abel$_a$ believes that if he$_a$ makes sure that Pip$_p$ has access to lots of money, he$_p$ will become one$_g$	(Output) Abel$_a$ will do his$_a$ best to make sure that Pip$_p$ has access to lots of money

One way we might portray the simulation alternative, is in terms of the belief, desire, intention, etc., boxes taken off-line. For clarity, I am going to talk in terms of pretend desire, pretend belief, and pretend intention, etc., boxes. According to simulationism, the above reasoning can be carried out by representations being tokened in various pretend boxes:

PRACTICAL REASONING PROCESS

PRETEND DESIRE BOX	PRETEND BELIEF BOX	PRETEND INTENTION BOX
(Input) Pip becomes a gentleman	(Input) If Pip has access to lots of money, he will become a gentleman	(Output) Make sure that Pip has access to lots of money

Here, the simulator imaginatively identifies with Abel, and engages her practical reasoning system to carry out the relevant reasoning.[19] I assume that it is in virtue of taking pretend input, that the reasoning system operates off-line.

Folk psychological reasoning can be quite complex. For example, someone might try to figure out what someone else thinks that a third party will think, do, feel, etc. Here, the representations involved in the reasoning become more complex. However, it is always the case that the Theory Theory requires more complex representations to be reasoned over than the Simulation Theory. In the more complex case just mentioned, Theory Theory posits the use of metameta-representations, and the Simulation Theory need, at most, posit reasoning in terms of metarepresentations. Thus, the following piece of reasoning:

> Hanna believes that Eric believes that if it's raining the streets are wet. Hanna also believes that Eric believes that it is raining. So, Hanna will come to believe that Eric believes that the streets are wet.

will look like this on a Theory Theory account:

REASONING PROCESS

BELIEF BOX	BELIEF BOX	BELIEF BOX
(Input) Hanna believes that Eric believes that if it is raining, the streets are wet	(Input) Hanna believes that Eric believes that it is raining	(Output) Hanna will come to believe that Eric will come to believe that the streets are wet

[19] Nothing should be made out of me talking about a reasoning *mechanism* and a practical reasoning *system.*

~~I first make adjus~~

~~I first~~

I imagine being Hanna pretending to be Eric.

~~Key~~ Nested simulation.

We do this sort of thing when reading fiction which has a narrator.

Simulationists can deal with the situation in one of two ways. Either they accept that simulators reason in terms of metarepresentations, or they claim that some kind of decoupling takes place after which the reasoning will be over object-level representations. According to the first option, the simulator imaginatively identifies with Hanna. She then reasons as follows:

REASONING PROCESS

PRETEND BELIEF BOX	PRETEND BELIEF BOX	PRETEND BELIEF BOX
(Input) Eric believes that if it is raining, the streets are wet	(Input) Eric believes that it is raining	(Output) Eric believes that the streets are wet

Notice that here the decision making system isn't deployed. It is only the pretend input and output that distinguishes this from a Theory Theory account of reasoning about Eric's beliefs. The representational complexity is compatible with a Theory Theory account at the basic level (reasoning about others). This form of simulation must, of course, both start and end with representations of the same complexity that Theory Theory posits in the relevant case, as the simulator must keep track of whose mental states are being simulated (Hanna's). So, in the case immediately above, the simulation starts with metametarepresentations, a sort of decoupling takes place, whereafter reasoning is carried out in pretend mode. At the end, the decoupled part of the metarepresentation is coupled with the outcome of the pretend reasoning process. The very beginning and end of a process of reasoning leading to a psychological attribution, are always the same on both theories.

The second option is to suppose that more decoupling takes place in the case of simulating Hanna. Hence, not only is 'Hanna believes' decoupled, but also 'Eric believes', and instead of replacing 'Hanna believes' with a pretend belief, it is 'Eric believes' that is replaced with such a belief:

REASONING PROCESS

PRETEND BELIEF BOX	PRETEND BELIEF BOX	PRETEND BELIEF BOX
(Input) If it is raining, the streets are wet	(Input) It is raining	(Output) The streets are wet

After this piece of reasoning, the pretend belief that is the conclusion loses its pretend mode and is combined with the decoupled parts of the initial represenatations. First the conclusion is attributed to Eric, with whom the simulator is identifying. On the assumption that Hanna would reason as Eric (another use of the assumption of similarity?), the conclusion 'Eric believes that the streets are wet', is attributed to Hanna as a belief. This way of dealing with reasoning about people reasoning about others, might be preferable to the one first mentioned because the actual simulating process does not involve metarepresentations. Whichever option the simulationist prefers, it remains the case that Theory Theory posits represenations that are more complex, for any given case of reasoning, than the Simulation Theory does.

One way of putting the difference between the two theories is to say that the Simulation Theory maintains that folk psychological reasoning can be carried out in terms of representations less complex than those posited by the Theory Theory to explain the same piece of reasoning. A folk psychological theorist must always reason minimally in terms of metarepresentations, but a simulator can (sometimes) reason in terms of simple representations.

Another difference between the two theories seems to be that, in the case of the Theory Theory, the reasoning process uses representations in the belief box as input, and the outcome of the reasoning is a representation tokened in the belief box also. In the case of the Simulation Theory, the immediate input to the reasoning are representations tokened in pretend boxes and the immediate output is tokened in a pretend box also. As I have presented matters, the processing mechanism need not be different in simulation and in theorising. Traditionally, simulationists have insisted that when an

A bit muddled. Simulation of behaviour (ie decision) involves the DM; Heal is pointing to simulation of reasoning involving the theoretical reasoner.

action is being simulated, the processing is carried out by the decision making or practical reasoning system. However, Heal's version of simulationism would not fit such a picture. This is due to the fact that she takes simulation of discursive reasoning as being one of the central cases of ?simulating reasoning. In such a simulation, it is not the decision making system that is deployed, but some theoretical or discursive reasoning system(s). Also, there is no particular reason that the simulationists shouldn't want simulation to take care of prediction and explanation of thought processes, and such simulation cannot take place in the decision making system.

In sum, it is unnecessarily restrictive to the Simulation Theory to assume that simulation must always be realised by psychological processes in the decision making system. Therefore, the differences between the Theory Theory and the Simulation Theory boil down to the following two:

ST doesn't.

- According to the Theory Theory, folk psychological reasoning minimally involves transitions among metarepresentations. According to the Simulation Theory, the transitions can be among simple, or object-level, representations.

see prev. remark

- According to Theory Theory, it is the belief box that produces the immediate input and receives the immediate output of folk psychological reasoning. According to the Simulation Theory, this function is carried out by pretence boxes.

good.

Davies seems to regard the most important part of the minimal distinction as being the difference in representational complexity. If simulation works with object-level representations, that exempts it from being treated as tacit knowledge of folk psychological theory (1994, p. 117). What I take this to mean is that for a simulation process to be indistinguishable from a reasoning process deploying tacit knowledge, this process must involve metarepresentations. In the simple case, simulation doesn't. However, a simulation might involve metarepresentations; when we reason about what Hanna will believe Eric will believe, for example. This process of reasoning is indistinguishable from a tacit knowledge process. This need not be a problem if one develops Davies' view in the way I have just done. What we need in order to provide a principled distinction between

No

practical v theoretical reasoning.

simulationism and Theory Theory is a comparative analysis of the reasoning processes involved on either theory. The thing is that in the case of Hanna's thoughts about Eric, Theory Theory should posit processes that involve metametarepresentations compared to the metarepresentations or object-level representations posited by the Simulation Theory. It is this fact that will distinguish the two approaches.

I must admit to not being entirely certain about Davies' idea. It seems to me that the account of tacit knowledge that he presents, doesn't naturally lend itself to a formulation in terms of the representational complexity that we require for the minimal distinction. As I understand it, an account of a derivational structure mirroring a causal structure will be underdetermined, and it is not Why? clear that it can offer a distinction so fine-grained as that between representations, metarepresentations, metametarepresentations, and say more so on. It is then not entirely clear that the threat of collapse introduced by a particular tacit knowledge construal of Theory Theory can be warded off by way of the minimal distinction unless the representation~~al~~s involved are taken seriously. That is, the representations will not simply be posited because there is a mapping between a certain causal structure and a derivational structure, but because we assume there actually to be representational states of varying complexity. We need a separate reason for this assumption, but I cannot go into that here. To conclude, I propose to take the minimal distinction as involving representational realism.[20]

Lastly, let me just address an issue about the complexity of Theory Theory. In more complex cases of folk psychological reasoning, what does Theory Theory claim we do? That is, in the case of working out what Hanna will believe that Eric will believe, do we need to consult a generalisation about what people think that other people will think? I don't think this is necessary. It is most plausible that we have a generalisation to the effect that people believe the same theory that we do - namely folk psychological theory, and that they deploy this in figuring out what people will do. I then simply harness whatever

[20] There are, of course, other options. One might reject that folk psychological knowledge is tacit. I shall discuss this issue in chapter 5. Alternatively, one can opt for another way of drawing up the debate altogether. Heal (1994b) rejects approaches at what she calls the "sub-personal level" (p. 132). The minimal distinction is such an approach. Instead, she attempts a distinction at the level of the person in terms of abilities or capacities.

generalisations are appropriate for working out what Eric will do on the assumption that Hanna would do the same, and hence arrive at the same result. On the face of it, this may seem similar to the kind of decoupling that takes place in a simulation of the same sort. However, there is no decoupling on the Theory Theory account, simply a number of generalisations being deployed in the reasoning (for example, since Hanna believes what I believe, she believes that if *a* desires that *q*, and so on and forth). So, the minimal distinction between the Theory Theory and the Simulation Theory keeps the threat of collapse at bay.

5. Rationality and Simulation

Before we can leave the threat of collapse behind us, we need to establish that the minimal distinction sets up each theory in a fashion that is congenial to what the various proponents have claimed. There can be little doubt that it appropriately captures the commitments of Theory Theory. When I reason using a theory about how psychological properties interrelated, it is only natural that the contents of my mental states will be metarepresentations - the psychological properties will be explicitly represented. This, of course, is not necessary on the Simulation approach. The minimal distinction appears very congenial to this approach also. It doesn't seem quite right to capture the idea of imaginative identification in terms of me imagining that 'I believe that if it is raining, the streets are wet, and I believe that the streets are wet', Rather, I imagine that 'if it is raining, the street are wet, and it is raining'. From which I imaginatively conclude that 'the streets are wet' *not* 'I believe that the streets are wet' (this comes later in the sequence that leads up to the attribution of the relevant state or action to the agent that is being simulated).

It seems, then, that the minimal distinction saves us from the threat of collapse whilst not reconstructing the competing theories in uncongenial ways. However, on closer scrutiny, the minimal distinction is not entirely unproblematic. For there is a version of the Simulation Theory - Rational Simulationism (Davies & Stone, 1998) - that, if we try to classify it according to the minimal distinction, falls neither on the side of the Theory Theory, nor on the side of the

Simulation Theory. This appears to render the distinction otiose and brings the debate back to the brink of collapse.

In order to examine Rational Simulationism in detail, we need to take a somewhat lengthy detour *via* Heal's Simulation Theory. The reason is that Rational Simulationism is really just a reformulation of ideas found in Heal (1996 & 1998). In fact, Davies & Stone take Rational Simulationism as saving Heal's basic ideas by presenting them in a different format.

The central tenet of Heal's idea is that there are norms of right reasoning - even means-end reasoning. It is adherence to such norms, rather than some form of semi-automatic imaginative process, that enables us to simulate. This proposal departs from more traditional accounts because the assumption of similarity is rephrased as an assumption of rationality or intelligibility. Rather than the simulator proceeding to simulate along the lines of 'what would I do under these circumstances', the simulator proceeds by asking herself 'what is the right thing to think, do, intend, feel, and so on, under these circumstances'. The norms that an agent recognises in her own reasoning are what enable her to simulate, on the assumption that people are rational or intelligible, and it is assumed that what makes them so is (largely) their adherence to the same norms. So, since the same norms of reasoning guide our thinking and acting, all we need to do to understand others, is to deploy these norms imaginatively. However, not all thought and behaviour is guided by these norms. Therefore, simulating is restricted to the following prime cases: (Heal, 1996, p. 56)

> The kind of simulationism I would like to defend says that the only cases that a simulationist should confidently claim are those where (a) the starting point is an item or collections of items with content, (b) the outcome is a further item with content, and (c) the latter content is rationally or intelligibly linked to that of the earlier item(s).

This doesn't quite mean that other cases of folk psychological attribution are ruled out. Prediction of action and emotional response[21] can also - to some extent - be explained by the Simulation Theory. However, the central cases concern the transition between

[21] Why Heal thinks that emotions are not really contentful states is unclear.

contentful psychological states and are constrained by some notion of rationality.

Norms of right reasoning do not simply boil down to the rules of logic, probability theory, decision theory, and the like. More links than those licensed by these disciplines are rational links in Heal's sense. Rules or principles from such disciplines may be contained in the norms, but they are neither exhaustive of such norms, nor are all such rules part of the norms: (p. 57-8)

> the simulation approach [...] recognizes that people do their reasoning, form their stances and take their decisions in real time, often under pressure, and facing the need to handle a great amount of complex material. [...] Hence not everything 'irrational' in the strict sense falls outside the domain of simulation. For example, being taken in by fallacious reasoning is something we can often sympathize with, find intelligible and predict by simulationist methods. The important issue for the applicability of simulation is whether we can see what went on as the upshot of the exercise of cognitive skills, not whether it was a flawless exercise of those skills. It is a corollary of this that intelligibility is not an all or nothing matter.

Another reason that 'rational' should not be taken in a strict sense, is that Heal wants to allow for simulation of utterances, emotions, and expressive behaviour, such as hugging someone, making angry gestures, and so on. In these cases, it is not so much the rationality of the production of such states that make them intelligible, but, more loosely, "the fact that we can often see 'from the inside' so to speak, why such actions are done." (p. 58).

'Rational', then does not well capture this approach to reasoning, but rather such expressions as "'such that some intelligible sense or point can be seen in it' or 'such that some justificatory account of it can be given'" (p. 58). 'Intelligibility' - or being able to see something 'from the inside', for that matter - is a *much* broader notion than 'rationality'. It covers all that is rational and a great deal more besides. Another difference is that 'being rational' is an intrinsic property. 'Being intelligible', on the other hand, is a relational property. This means that for something to be intelligible requires something from both ends of the relation. A person's intelligibility does not simply depend on how she reasons - whether she is rational, say - but also on the intelligence of the person that tries to understand her. So, it may be that what is intelligible to one person is unintelligible to another. The possibility of a simulative understanding depends on the agents being suitably related to each other.

But now we seem to be going around in circles. Simulation only applies to intelligible thought and behaviour because simulators do not use their own reasoning as a measure, but *correct* reasoning. The problem is that Heal then extends what counts as correct reasoning to reasoning that we 'can see from the inside' or make some sense of. However, it is hard to see what that sense can be, other than 'intelligibility', since we agree that talking of *correct* reasoning doesn't apply in a number of cases that we can simulate. If this is what Heal says, then all she is saying is that we can understand thought and behaviour because it is intelligible. But the intelligibility of thought and action is what we are trying to get at. So, Heal's position seems to be of little help here.

Even if we assume that there is some non-circular interpretation of Heal's position, it is still deeply problematic. The problem is that 'intelligibility' is too loose a notion to play the role that Heal wants it to play. To see this consider this concrete example of action that, according to Heal is unintelligible.

The Langer effect is named after the discoverer of the effect, Ellen Langer (Langer, 1975)[22]. In her experiment, subjects are given lottery tickets as a reward for their participation in some prior psychological experiment. Some are simply presented with tickets, whereas others are allowed to choose which of a number of tickets they want. When the experimenter subsequently asks the subjects to sell her back their tickets, it turns out that the subjects who chose their own tickets demand a higher price than those who weren't given any choice. When other subjects are asked to predict the behaviour of the original subjects, they fail to take this into account, but predict that all the subjects will ask roughly the same price for their tickets.

As Heal points out, there have been problems replicating this experiment (Kühberger et al., 1995). However, for the sake of argument

[22] What I go on to describe is not the original set-up of the experiment. Rather, I follow Nichols et al.'s reconstruction of it, since it is the failure of other subjects to predict the behaviour of the experimental subjects that is at issue here (Nichols, Stich, Leslie & Klein, 1996).

she assumes that it is a genuine effect. She also takes it to be a prime example of irrational behaviour, and consequently a case that her brand of simulationism cannot explain. Assuming with her that the effect is genuine, we ask ourselves is it really unintelligible?

The answer seems to be yes and no. It is no because a good case can be made for the fact that one may be more attached to a lottery ticket that one has chosen oneself, than one that one has been given. One often chooses a ticket that means something to one - that has the initials of a loved one, that has a number that has some special significance, such as one's birthday (see Langer's own observations). These features are imagined to be lucky features. Consequently, one thinks that one has more chance of winning the lottery with *this* ticket rather than another one. If one is simply given one, it may be that it possesses no lucky features that one can think of - the numbers or letters are wrong. Or it may simply be the case that one has not endowed it with such features and finds it difficult to do so in a no-choice situation. All this makes the behaviour of the subjects quite intelligible, in particular, if one takes on board Heal's idea of seeing it 'from the inside'. I know that when I choose a lottery ticket, I look for a special feature that I connect with an increased chance of that ticket winning the lottery. I may also know that this belief is false, but that doesn't make the behaviour unintelligible. In another sense of 'intelligible' - a more narrow rational sense - the Langer effect does not make sense. There is no such thing as a lucky ticket, and presumably the failure to predict the Langer effect is due to the fact that we are all aware of this. According to probability theory, all tickets are equally likely to win, so subjects should all sell their tickets at roughly the same price.

The fact that we fail to predict the Langer effect (Nichols, Stich, Leslie & Klein, 1996) seems to indicate that the subjects' behaviour doesn't make sense. We are probably surprised once we learn of the effect. The Langer effect is not immediately obvious or intelligible. But if one starts to meditate on the various factors that may be involved, it begins to become more and more intelligible. And Heal is emphatic about including the results of such meditation under the intelligible or rational (1996, p. 58).

The above makes it abundantly clear that 'intelligibility' is so loose a notion that one can reasonably say of the very same action or thought that it is both intelligible and unintelligible (albeit not in the same respects). This does not simply seem to be a question about intelligibility being a matter of degree, or the case not being clear-cut. The problem is that the notion of 'intelligibility' is so loose as to fit almost any relation between psychological contents.[23] The reason this is the case, is that the intelligibility or rationality of people is more properly understood as forming a background assumption to any psychological theorising. Or, as Donald Davidson, would put it, a constitutive ideal: (1970, pp. 222-23)

> The point is rather that when we use the concepts of belief, desire, and the rest, we must stand prepared, as the evidence accumulates, to adjust our theory in the light of considerations of overall cogency: the constitutive ideal of rationality partly controls each phase in the evolution of what must be an evolving theory.

The theory that Davidson refers to is psychological theory. So, it is only on the assumption that people are rational that we can practice psychology at all. But playing a constitutive role in the background is a far cry from playing an essential role in the foreground. *Any* psychological understanding requires some form of rationality on the

[23] Stich & Nichols (1997) complain that Heal has made her theory unfalsifiable. Heal's paper is written in response to a particular line of approach taken by Stich et al. (Nichols, Stich. Leslie & Klein, 1996; Stich & Nichols, 1992, 1995, 1997). Stich, Nichols, Leslie & Klein have pressed a notion of cognitive penetrability to serve as a dividing line between Folk Theory Theory and Simulation Theory. It is supposed to work as follows. Since, when we simulate, we just use whatever mechanisms we use when we reason ourselves, information about how people reason should play no role in a simulation. On the other hand, in Theory Theory such information does play a role. One way to put this, is to say that a simulation is cognitively impenetrable - it is immune to information about how people reason. Experiments can then be set up to determine whether the Theory Theory or the Simulation Theory is correct. The Langer experiment has been a key case taken to support the Theory Theory. The reason is that if psychological prediction were due to simulation, the Langer effect should be replicated by the subjects asked to predict the sell-back price of the experimental subjects, since the effect would be hard-wired in decision making. However, theory theorists can claim that it is due to lack of information that subjects fail to predict the Langer effect - our folk psychological theory is not complete.

Heal (1996) has pointed out a number of shortcomings with this view (whereas simulation cannot be affected by *absence* of information, it can be influenced by *presence* of information about decision making and reasoning procedures), as well as problems with the actual example. Her simulationism, however, is supposed to rule out that simulation can help us understand cases like the Langer effect because such an effect is irrational or unintelligible. Given that this is part of the purpose of her theory, it is a serious shortcoming that what constrains what the Simulation Theory can explain is so loose that almost any counterevidence can be accounted for. In short, Stich & Nichols conclude, rightly I think, that Heal has immunised her Simulation Theory to falsification through counterevidence.

part of the subject, including scientific psychology and psychoanalysis. This means that it is hard to see how rationality or intelligibility, on its own, can be harnessed to play a role in specific predictions and explanations. What we need is an explanation of *how* something makes sense in the light of something else. And Heal's account does not seem able to provide us with this.

It is instructive, at this point, to see how Theory Theory deals with rationality as a constitutive ideal. We might look at Davidson. He talks of a "common-sense scheme for describing and explaining actions" (1974). He uses 'scheme' instead of 'theory' because he does not believe that we can have strict psychological laws - and the laws of science are strict, 'theory' being understood on the model of 'scientific theory'. But it is quite clear that here specific generalisations are what carry the weight of psychology, not the assumption of rationality. No theory theorist denies the constitutive role rationality plays in psychology, nor do they deny that psychological explanation is also rational explanation.[24] The point is simply that we need something more specific than an assumption of rationality to do psychology. We need specific correlations. These correlations are described in folk psychological theory.

Having been introduced to the idea of rationality or intelligibility as playing a key role in simulation, we can now return to Davies & Stone's Rational Simulationism to see how it clashes with the minimal distinction.

6. Rational Simulationism and the Minimal Distinction

Rational Simulationism is an attempt to save Heal's thesis through a reworking of it. In fact, it saves it by providing the specifics that are necessary for it to be workable. So, whereas the assumption of rationality is accepted as a background assumption, the foreground is taken up by a *theory* of right reasoning. When we simulate, we pretend that we are someone else and decide what to think or do on the basis

[24] If, however, we accept that results from scientific psychology can be integrated into folk psychological theory, we open the possibility of certain folk psychological generalisations being less than rational. Nevertheless, it still seems to be the case that any psychological theory relies on people being rational or intelligible to some extent.

of our theory of right reasoning, on the assumption that the agent that we are simulating is rational.

Our theory of right reasoning may be wrong. We may have false information and there may be information about how best to reason that is not (yet) incorporated into our theory. Thus, it is not simply the case that we cannot simulate thought or action that is irrational in the sense of not falling under a principle of right reasoning; there may even be rational thought or action that we cannot simulate because our theory is false or incomplete. Failure to predict the Langer effect, for example, can be explained either by this effect constituting genuine irrational behaviour, or by it constituting behaviour that is rational but not included in our theory.[25]

The problem Rational Simulationism faces is that it seems to fall *between* the Simulation Theory and the Theory Theory, since it holds that folk psychological reasoning must minimally involve metarepresentations, but also that it is a case of imaginative identification. If we have a theory of right reasoning, that theory must contain information about psychological categories. If it didn't, it wouldn't be a *theory* of right reasoning. It could be a theory of logical implication, of probability, or the like. But it is a theory of right reasoning that Davies & Stone attribute to subjects. I quote Stein on this issue: (Stein, 1996, p. 5-6)

> Rules of logic apply to statements and determine the logical relations among them; principles of reasoning that stem from rules of logic apply to beliefs and determine the relations among them. Some, but certainly not all, principles of reasoning are based on rules of logic. According to the standard picture of rationality, principles of reasoning based on rules of logic are normative principles of reasoning. As another example, consider the following rule of logic:
>
> MODUS PONENS: *A* and **if** *A*, **then** *B* together entail *B*.
>
> This gives rise to the following normative principle of reasoning:

[25] This should also answer Stich & Nichols' complaint about unfalsifiability. We can only successfully simulate thought and behaviour that can be understood in terms of our theory of right reasoning

> MODUS PONENS PRINCIPLE: if you believe *A* and you believe **if** *A*, **then** *B*, you should believe *B*.

If this is right, it means that the representations employed in reasoning leading to psychological attribution must minimally be metarepresentational on the rational simulationist view. If your folk psychological reasoning is shaped by deployment of generalisations like the above modus ponens principle, the immediate input and output of the simulation must be metarepresentations. The modus ponens principle cannot operate on simple representations, like *A*, because it does not apply to such representations, only to metarepresentations, like the belief that *A*.

Secondly, Davies & Stone maintain that in Rational Simulationism there is imaginative identification, and hence the immediate input and output of a simulation are pretence box ? representations. This combination of viewpoints appears to place Rational Simulationism right in the middle between Simulation Theory and Theory Theory on the minimal distinction. However, rather than this being a shortcoming, I take it as being a virtue of the distinction. Intuitively, holding that we have a *theory* of right reasoning makes Rational Simulationism a Theory Theory, whilst maintaining that we imaginatively identify with subjects is traditional simulationism. The minimal distinction tracks these intuitions pretty precisely.

It would seem that since there are two elements to the minimal distinction and these two elements can come apart in theorising, that some theories will fall between Simulation Theory and Theory Theory, not properly being either. This, by itself, is not a problem. However, I think that Rational Simulationism is best seen as a version of the Theory Theory. The reason is that imaginative identification, which is what makes Rational Simulationism simulationist in character, is redundant.

7. Rational Simulationism as a Theory Theory

In traditional Simulation Theories, the role of imaginative identification is to allow you to harness abilities and capacities that you deploy in your own reasoning and decisions, that you need not be aware of. The point, of course, is that once you are aware of what this

speed
efficiency.

capacity consists in, there is no reason for you to imaginatively identify ? with subjects. You might as well just apply the relevant principles directly to the subject. To put it differently, a theory of right reasoning quantifies over how rational agents will think and act. As such, it is applicable to all metarepresentations - whether the subject is represented as me or as someone else. This means that imaginatively identifying with a subject becomes superfluous.

It is perfectly consistent with the main tenet of Rational Simulationism that we discursively or theoretically reason when we attribute psychological states to subjects, and hence that the immediate output and input of the relevant reasoning processes are those of the belief box. It is not just that nothing about possessing and deploying a theory of right reasoning requires us to simulate subjects, it is much stronger than that. Once you assume that knowledge of a theory is causally efficacious in the production of folk psychological attributions, it makes no sense to require that subjects imaginatively identify with the subjects that they want to understand. It makes no sense because it is pointless. Therefore, I think Rational Simulationism is best regarded as a version of the Theory Theory.

Davies & Stone are, of course, adamant that the claim that we have a *theory* of right reasoning does not amount to giving up simulationism and embracing Theory Theory. According to them, there is an important difference between a Rational Theory Theory and Rational Simulationism: (Davies & Stone, 1998, p. 61)

> the simulation theory is clearly not proposing that we make predictions by the disengaged use of a set of normative principles about reasoning. Rather, normative principles may be used in simulation because they are already available to us when we ourselves engage in reasoning.

They take the notion of critical reasoning from Tyler Burge (1996). According to Burge, critical reasoners are reasoners that are reflectively aware of the activity of reasoning - can evaluate it as being good or bad reasoning. Being critical reasoners is "an essential part of normal adult reasoning as we know it" (Davies & Stone, 1998, p. 61). By an *engaged* use of normative principles, I take it that Davies & Stone mean something similar to what traditional simulationists

?

mean. Let us remind ourselves of that. Traditional Simulation Theory has it that it is constitutive of your being able to understand other agents that you are an agent yourself. Otherwise you wouldn't have the decision making and hypothetical reasoning procedures to deploy in a simulation. The 'engaged' here refers to the fact that you are harnessing some ability that is essential to you as an agent, to play a role different from its usual one. Similarly, Rational Simulationism seems to say that you need to have a theory of right reasoning in order for you to be a critical reasoner. And being a critical reasoner is essential to you as an agent.

Not really.

Compare both Traditional and Rational Simulationism with Theory Theory. Is it constitutive of your understanding of others that you are an agent or a critical reasoner yourself? Presumably not. It seems possible that an alien with a different psychology to ours might acquire folk psychological theory. As long, of course, as she is rational. It is hard to see how an irrational being could acquire a theory. However, if she is rational, she should not only be able to acquire a theory of right reasoning, but also to use it in her own case. Indeed, it seems that if she were rational, once the theory of rational reasoning is acquired, she *would* use in her own case, since that is the rational thing to do. However, using such a theory in one's own reasoning just *is* being a critical reasoner.[26] So, accepting Davies & Stone's idea that it is a theory of right reasoning that makes you a critical reasoner, any alien that were to acquire such a theory would become a critical reasoner. Then it seems that being a folk psychologist and being a critical reasoner go hand in hand. Both are by-products of acquiring a theory of right reasoning. The engagedness of Rational Simulationism is connected with the acquisition of the theory of right reasoning. Whereas this is clearly different from Theory Theory as I have stated it, it does not seem incompatible with it.

?

The similarity between Theory Theory and Rational Simulationism will then be that any rational agent would be able to acquire either. The difference is, that acquiring a theory of right

[26] If the alien were already a critical reasoner, having some idea or theory of right reasoning, then she would either possess roughly the same theory as us or a different one. Acquiring our theory of right reasoning would lead to remodelling overall - of ours, hers, or both. But this is just the situation we are in when we come to realise new things about right reasoning. Having a theory of right reasoning will be constitutive of her being a critical reasoner. Having *our* theory of right reasoning will - insofar as it is different and right - be at least partly constitutive of her being such a reasoner. This seems to fit nicely with Burge and Davies & Stone.

reasoning will turn you into a critical reasoner if you are properly rational; acquiring folk psychological theory won't. This fact, though, is an artefact of what a theory of right reasoning is *about*, *not* the fact that it is a theory. It seems to me, then, that the attractiveness of reclassifying Rational Simulationism as Rational Theory Theory remains. The reason is that folk psychological reasoning minimally deploys metarepresentations, imaginative identification is not required, and the engagedness of the use of a theory of right reasoning is an artefact of its content, not the fact that it is a theory.

Rational Theory Theory *is* different from the standard form of Theory Theory presented so far. According to it, we don't reason along the lines "people do, think, etc., this and that under these circumstances", but "it would be right for people to do, think, etc., this and that under these circumstances". The main difference is that between 'ought' and 'will'. Where traditional Theory Theory has something like:

> (G) If *X* wants to Ø, and *X* believes that *A*-ing is a way for her to bring about Ø, then *X* will *A*, *ceteris paribus*.

Rational Simulationism has

If X wants to ϕ, and X believes that A-ing is the right / best way to Ø, then X will A.

?

> (N) If *X* wants to Ø, and *X* believes that *A*-ing is a way for her to bring about Ø, then *X* ought to *A*, *ceteris paribus*.

I use '(N)' for norm as opposed to '(G)' for generalisation. The difference between the two accounts appears to amount to agents being able use (N) directly in deciding what to do. It is much more unlikely that (G) could be used so. It seems odd that I should decide what to do on the basis that this is what agents generally do under these circumstances. However, given that the background of the generalisation is rationality, that is that the agents that act in this fashion are rational, there is perhaps some role it can play. For example, rational agents generally try to optimise in their decision making. If I want to optimise also, then (G) can be taken to heart. However, there is little doubt that (N) is more straightforwardly applicable in our own reasoning. (G) does not lend itself to use in decision making in the same way. Having said all this, I shall leave Rational Theory Theory behind. What follows in the next chapters should be applicable to all versions of Theory Theory, Rational Theory Theory included.

I have argued that Rational Simulationism does not provide a threat to the minimal distinction. In the one case where the distinction classifies a theory otherwise than it classifies itself, this has been due to the fact that the theory vacillates between the Simulation Theory and the Theory Theory. Rather than regarding it as neither of the two, I have argued that it is best seen as a Theory Theory. This means that we can stave off the collapse of the Theory Theory versus Simulation Theory debate.

8. Is Theory Theory a Folk Theory Theory?

Now that we have satisfied ourselves that the minimal distinction serves to uphold the Theory Theory versus Simulation Theory debate, we can conclude that Folk Theory Theory is a tenable option. According to it, we have knowledge of a folk psychological *theory* because we have knowledge of some body of information the usage of which in reasoning minimally involves transitions among metarepresentations.

However, as I have already indicated, I find the Folk Theory Theory position unsatisfactory. The reason is quite simply that it lets too much be a theory and, consequently, too much be a Theory Theory. There is a loose use of the term 'theory' that lends itself to Stich & Nichol's understanding of it. We were introduced to it, as well as to a stricter one, at the beginning of this chapter. It would, perhaps, be churlish to insist that only the stricter version is correct. However, we can safely maintain that it might be better using the term in its stricter sense for purposes of precision and informativeness. This will have the positive consequence that when you say that something is a theory, I will have a much better idea of what you mean than if you were to use the term in its looser sense. This is not linguistic fascism - if anything, it is linguistic parsimony; if we agree on a relatively tight and precise meaning to our words, we don't need to spend hours discussing just what nuance of the term we have in mind (although we probably won't quite be able to avoid this).

We could avoid debate
by agreeing to the tighter or
the looser sense. What
matters is that the ambiguity
is resolved

Many theorists about folk psychological/physical/biological knowledge regard it as being importantly different from that of a theory - its generalisations are too loose, there is no coherent structure, etc. They would be justified in complaining that Stich & Nichols' deflationary understanding of 'theory' lumps their theory of folk psychology together with views that involve a much less distinctive understanding of 'theory'. It seems highly misleading to regard this class of theories as amounting to more or less the same position. Fairly substantial differences between the proponents of Folk Theory Theory would be allowed. I side with these protestants. Using 'Theory Theory' to cover such a heterogeneous class of positions is very unhelpful. We're in enough trouble as it is, theory theorists differing on a number of other issues. If we also understand 'theory' in terms of just any internally represented body of knowledge, pandemonium will ensue. Much too much could be meant by 'Theory Theory'.

In conclusion, we are looking for a tighter notion of 'theory' to characterise the claim of the Theory Theory. We want something relatively specific to be meant by it, be it Theory Theory of folk psychology, folk physics, or folk biology. This makes for more distinctive and more falsifiable theories. And, as I shall argue in the next chapter, we don't have to look far for such a notion.

such as?

Chapter 3

Folk Psychology & Scientific Theory

We understand others, as well as we do, because we share a tacit command of an integrated body of lore concerning the lawlike relations holding among external circumstances, internal states, and overt behavior. Given its nature and functions, this body of lore may quite aptly be called "folk psychology". (Churchland, 1981, p. 256)

All these characteristics of theories ought also to apply to children's understanding of mind [...] such theories should involve appeal to abstract unobservable entities, with coherent relations among them. Theories should invoke characteristic explanations phrased in terms of these abstract entities and laws. They would also lead to characteristic patterns of predictions, including extensions to new types of evidence and false predictions, not just to more empirically accurate prediction. Finally, theories should lead to distinctive interpretations of evidence; a child with one theory should interpret even fundamental facts and experiences differently than a child with a different theory. (Gopnik & Wellman, 1992, p. 234-5)

On the more narrow reading of 'theory', scientific theories are paradigm cases of theories (Carey, 1985; Churchland, 1981; Gopnik & Meltzoff, 1997; Gopnik & Wellman, 1992 & 1994; Wellman, 1990). Any body of information that wants to call itself a theory must therefore share some features with scientific theories. On this narrow reading, then, the 'theory' of folk psychological theory is modelled on the 'theory' of scientific theory, like the theories of physics, biology, geology, and physiology.

In comparing the body of folk psychological information to scientific theories, theory theorists point to a number of features that scientific theories supposedly share. In choosing these features, theory theorists are inspired by the philosophy of science. Models of scientific theories by, among others, Carl Hempel, Ernst Nagel, and Thomas Kuhn, play a large part in modelling the theoreticity of folk psychology (Hempel, 1965; Kuhn, 1970a; Nagel, 1961). However, this does not mean that theory theorists believe that the body of folk psychological information is a *scientific* theory as it stands. To my knowledge, no one does. There is, however, considerable discussion about whether folk psychological theory is sufficiently *scientific* to be able to form the basis of a scientific psychology - that is, whether it is a good (scientific) theory (yes: Fodor, 1987; Horgan & Woodward, 1990; no: Churchland, 1981; Stich, 1983).

Different point

Theory theorists, then, are interested in what makes a theory a theory. The model is scientific theories, but obviously not all theories are scientific theories. So, only some of the features possessed by scientific theories are picked out as being sufficient for something being a theory. One way to divide this up would be to claim that the static features of scientific theories make them theories, and the dynamic features make them scientific. The static features would capture the structure of a theory, the dynamic ones how it comes about, how it is tested, and so on. *Scientific* theories are, for example, subjected to rigorous experimental testing. However, most of the theory theorists that defend this modelling of 'theory', are also eager to show how the dynamic features of the body of folk psychological

astronomy?

information are similar to those of scientific theories. Children's developing understanding is importantly similar to scientists' changing understanding of the world. Since my focus is Synchronic Theory Theory, I shall ignore, whenever possible, dynamic features of theories. Suffice it to say that even where it is assumed that the model of 'theory' includes dynamic features, it is still mostly accepted that there are interesting differences between folk psychology and science, although this might be more a matter of degree - more rigorous testing in the latter, and so on. Henceforth, when I talk of the features of scientific theories, I mean the features that make these bodies of information *theories*, not what makes them *scientific* theories.

In this chapter, I will examine the idea that the body of folk psychological information is a theory because it has certain features in common with scientific theories. I will proceed as follows. I will present two suggestions by theory theorists of how to model theoreticity. The two differ according to what ideas in the philosophy of science they stress. Each claims that scientific theories possess a number of features that any body of information must have in order to be considered a theory. There are, then, two parts to the argument: an assertion of what counts as a theory by reference to the nature of scientific theories, and an assertion that the body of folk psychological information is a theory according to this view. An evaluation of these claims must consequently fall in two parts. First, we must examine whether the construal of theoreticity is acceptable and, insofar as it is, we have to consider whether the body of folk psychological information possesses the relevant features. Only if both claims are true, can we accept the view. If the first is true and the second false, then we should conclude that the body of folk psychological information does not constitute a theory. If, on the other hand, the first is false, we can conclude nothing about the theoreticity of folk psychology.

It is important to be clear about what the guiding principles are. Theory theorists are looking for conditions of theoreticity that are jointly sufficient and, as far as possible, individually necessary *ceteris paribus*. It is not the aim of Theory Theory to classify the body of folk psychological information as a theory on the basis of the fact that it has similarities with a couple of freak scientific theories. That is, there may be bodies of information of a highly idiosyncratic structure having little, if anything, in common with the majority of scientific theories.

These may yet be regarded as scientific theories. For its claim of legitimacy, Theory Theory is looking for features generally shared by scientific theories. There is little hope coming up with a list of necessary and sufficient conditions. We are not simply interested in necessary conditions, so our aim is to find conditions that are jointly sufficient and individually close to being necessary (to rule out the freak cases). I propose to go about this project in the following way. As a rule of thumb, I will look for characteristics that as many scientific theories as possible have in common. If a suggested condition of theoreticity is such that a number of scientific theories don't possess it, I will reject it.

I will conclude that a reworking of a more traditional picture of scientific theories serves well as the model on which to base one's notion of 'theory'. A Kuhnian model is more problematic. Most importantly, it does not provide an account of what scientific theories have in common *qua* theories. If anything, it is rather to be considered an account of a particular kind of scientific theory. However, this is not what we are interested in at this point. What we want to know, is what makes the body of folk psychological information a theory in the first place.

1. Folk Psychology & Scientific Theories: the Traditional Approach

The oldest and most influential version of the idea that folk psychology is similar to scientific theories dates back to Wilfrid Sellars' seminal paper: "Empiricism and the Philosophy of Mind" (1963). Here, Sellars stresses the continuity between everyday thinking and scientific thinking: (p. 183)

> science is continuous with common sense, and the ways in which the scientist seeks to explain empirical phenomena are refinements of the ways in which plain men, however crudely and schematically, have attempted to understand their environment and their fellow men since the dawn of intelligence.

Note that 'being continuous' with is not 'being identical with', wherefore the possibility of there being important differences between everyday thinking and scientific thinking remains open. The body of folk psychological information is *like* a scientific theory, not literally a scientific theory (p. 183).

True to the period, Sellars prefers to analyse the notion of 'theory' in terms of the language of theories. Therefore, for him, the most important similarity between scientific theories and the body of folk psychological information, is the fact that psychological terms can be regarded as theoretical terms; terms that refer to unobserved entities. This allows him to show how our conception of ourselves and others as subjects of psychological states, is compatible with such states not being introspectively *given* to us. Rather, they may simply have been *posited* to explain why we do what we do. Indeed, given all the similarities between thought and speech, it is quite likely that speech has served as a model for the theory of thought and action. Thoughts are unuttered sentences. Just as sentences mean this or that, thoughts have meaning. Certain dissimilarities are allowed, of course: "the episodes in question are not the wagging of a hidden tongue, nor are any sounds produced by this 'inner speech'." (p. 187).

This general idea has been followed up more recently by Gopnik, Meltzoff, and Wellman (Gopnik & Wellman, 1992 & 1994; Gopnik & Meltzoff, 1997), all developmental psychologists. Their idea of what a scientific theory is, is very precise and clear-cut. I call this approach 'traditional', because it is largely inspired by a Hempelian-Nagelian idea of the nature of scientific theories (Hempel, 1965; Nagel, 1961):[27]

> A scientific theory:
>
> (i) forms a coherent whole in the sense that the theoretical terms are interdefined,
>
> (ii) postulates abstract entities or properties (referred to by the theoretical terms) that causally explain observables,
>
> (iii) contains postulates that are internally related in terms of laws,
>
> (iv) provides explanations phrased in terms of these abstract entities and laws (deductive-nomological explanations),
>
> (v) allows predictions not just within its own domain, but beyond it; i.e. the theory can be applied outside the domain it was originally meant to explain,

[27] What follows is a condensation of Gopnik & Wellman, 1992, pp. 233-35, Gopnik & Wellman, 1994, pp. 258-64, and Gopnik & Meltzoff, 1997, pp. 32-41. Cf. also the second introductory quotation to this chapter.

(vi) yields idiosyncratic interpretations. The interpretations must be specific to the theory in the sense that another theory of the domain must yield different interpretations. Descriptions of phenomena will not do, and

(vii) leads to false predictions.[28]

Since Gopnik, Meltzoff, and Wellman present us with such stringent criteria for theoreticity, it ought to be relatively straightforward to determine whether: i. scientific theories have the above characteristics, and ii. whether the body of folk psychological information has them. Below, I will examine each condition, starting with the *prima facie* most problematic ones; (v) and (vii).

(v). In the ordinary usage, the domain of a theory is that to which the theory applies. It is therefore a contradiction in terms to say that a theory should apply outside its domain. However, we can put this problem aside, if we understand the idea expressed by (v) to be that a theory should be applicable to properties and things, other than those to which the theory was originally constructed to apply. But even so, we face certain difficulties. There are at least two different ways of understanding the idea that a theory applies outside its original domain. Firstly, one might mean that a theory literally applies outside its original domain. In this sense, many central scientific theories clearly do not have such application. Newton's Laws of Motion apply exclusively to massive bodies. The principle of natural selection only applies to biological organisms.[29] We can therefore not accept this sense of the applicability of a scientific theory outside its (original) domain.

Not really

genetic programming?

A-Life systems?

Alternatively, one might suppose that a theory applies outside its own domain, when it is used as a model. For example, Maxwell used the idea of a universal medium of the propagation of light and heat as

[28] Gopnik, Meltzoff, and Wellman also posit dynamic features as characteristic of theories. Theories typically follow a particular course of development. The Copernican revolution is taken as the stereotype of such development. The ontogenesis of a theory of mind parallels such development (Gopnik & Wellman, 1992, pp. 235-239; Gopnik & Meltzoff, 1997, pp. 39-47). It would appear that the model of theoreticity that I will present in section 3, is taken to be what characterises the ontogenesis of folk psychological theory according to the authors here mentioned.

[29] Darwin, himself, talks of "beings", "organisms", and "life" in the context of natural selection (Darwin, 1872/1994). *The Cambridge Encyclopedia* stresses this even more: (p. 836) "**natural selection** The complex process by which the totality of environmental factors determine the non-random and differential reproduction of genetically different organisms..." (my underlining).

a model for his theory of the electromagnetic field (van Fraassen, 1980, p. 48). More recently, the astrophysicist Lee Smolin has used the idea of natural selection as a model for his theory of the cosmos. He argues that the universe and the laws of physics are subject to natural selection (Smolin, 1997). In neither of these cases is it strictly speaking true to say that the theory used as a model applies outside its original domain. The theory is changed in important respects to fit the new range of phenomena. Natural selection, for example, essentially involves reference to biological organisms. Whatever astrophysical phenomena are, they don't appear to be biological organisms, nor does Smolin suppose them to be so. Therefore, it seems that the best way to conceive of scientific modelling, is to say that what serves as a model, for example the theory of natural selection, does not literally apply outside its original domain, but certain features of it serve to highlight particular features of some other areas of theorising. The theory in which the old ideas are used as models, is best regarded as a separate new theory. As Sellars pointed out: "The essential thing about a model is that it is accompanied, so to speak, by a commentary which *qualifies* or *limits* - but not precisely nor in all respects - the analogy between the familiar objects and the entities which are being introduced by the theory" (1963, p. 182).

If we assume that it is the idea of being usable as a model that Gopnik, Meltzoff, and Wellman have in mind with (v), it seems rather a peculiar condition on theoreticity. In principle, presumably *any* body of knowledge with some coherent structure can be used as a model in theorising. Whether it is used in this manner is largely a matter of luck. It seems to be down to whether a new theory happens to be proposed that has certain interesting similarities with the old one, whether the proponent of the new theory was well versed in the old theory, whether the ideas of the old theory are clearer than those of the new one, and so on. In other words, extendibility as a model seems to be an accidental and external feature of scientific theories. I therefore conclude that we should not let the *theoreticity* of a body of knowledge depend on its extendibility as a model.

(vii) is a strange characteristic to require theories to have. Gopnik & Wellman flesh this out by saying that "theories are never completely right" (1992, p. 234). However, even if we have reason to be sceptical about whether we'll ever come up with a fully true theory, it should not form part of the characteristics of a theory that it cannot be true. Hence, I advocate skipping (vii) also.

I conclude that we should reject (v), and (vii) as forming part of the conditions for theoreticity. The reason is that, insofar as they form part of some sufficient conditions for being a scientific theory, they are very far from being necessary, since they are not met by a number of very prominent scientific theories. However, we are interested in conditions that are individually close to being necessary - not conditions that we know not to be met by many scientific theories.

?

Let us return to the other conditions, to see whether they fare any better. (vi) appears to be nothing but a function of (ii). If a theory posits abstract entities to account for the phenomena, then for two theories to count as different theories, they ought to posit different abstract entities. However, once you do that, you do not simply have a description of the phenomena, but already a distinctive interpretation of them. So (vi) is really only a corollary of (ii). Therefore, we can do without it. We are now left with a much shorter list of conditions (i) - (iv). Before being in a position to be able to embrace or reject this picture, we need to look a bit closer at what is meant by 'abstract entities or properties' and 'observables'.

No longer any ref. to prediction.

Talk of 'observables' has traditionally been connected with talk of 'nonobservables'. It was assumed that the language of theories contained observational terms that referred to observables and theoretical terms that referred to nonobservables (Hempel, 1965). More recently, observability has been defined thus: (van Fraassen, 1980, p. 16)

> X is observable if there are circumstances which are such that, if X is present to us under those circumstances, then we observe it.

This principle of observability is sometimes linked to the realism debate in the philosophy of science. For example, one might hold that only what is observable really exists.

*Non*observables, on the other hand, are supposed to be the entities or states of affairs that are referred to by theoretical terms. They are not observable, but play a causal-explanatory role *vis-à-vis* the observables. An example of a nonobservable entity would be a quark. Quarks are elementary particles. Their structure - along with the structure of leptons (other elementary particles) and certain binding elements: gluons - explains the structure of subatomic particles.

However, regarding observability as something absolute - either something is observable or it is not - is not congenial to the project of determining the nature of scientific theories. For too much is in principle observable, particularly if we allow observables by instruments. Oil drops, Neptune, atoms, and DNA are all observable. And not allowing observables by instruments seems unprincipled since what is detectable by the naked senses and what is observable by help of instruments, such as microscopes and telescopes, lie on a continuum. But if this is true, then (ii) is false, for Neptune, atoms, and DNA all play the role of abstract or theoretical entities in the relevant theories. We would therefore have observables explaining other observables. Scientific theories, however, have non-observables explaining observables. I don't think this is an unusual situation in the sciences at all. One can even make a case for black holes being observable. Black holes are supposed to explain observables, like the frantic activity of quasars. Quasars are thought to be the nuclei of galaxies in which there is much activity. This activity can be explained by the presence of a black hole, whose gravity draws matter in surrounding space into it. So, a black hole ought to play the role of the abstract entity in (ii). However, a good photograph of a quasar will show a black circle in the middle of the quasar that is the black hole around which the quasar revolves. So, black holes can be understood as observables.[30]

But there are reasons to believe that Gopnik, Meltzoff, and Wellman do not want to accept the traditional dichotomy between observables and nonobservables. They recognise that although the referents of theoretical terms are often nonobservable, some are

[30] That one cannot simply observe a black hole *as such* is not sufficient to show that it should not be regarded as an observable. The fact that one sees something and doesn't know what it is, doesn't make what is seen any less observable. Here one might distinguish between observing and observing that (cf. van Fraassen, 1980, p. 15). Sometimes one must know certain things about what one sees, or take surrounding factors into consideration. One must, for example, observe the movements of objects around it in order to ascertain whether anything is a black hole. In the same vein, think of how you might observe a philosopher. You cannot tell simply by looking. You need to see what they do.

observable, like genes.[31] This is why they talk of abstract entities and properties as opposed to nonobservable entities and properties. Abstract entities are "removed from, and underlying, the evidential phenomena themselves." (Gopnik & Wellman, 1992, p. 233). Unfortunately, this doesn't make the problem go away. First of all, to talk of the referents of theoretical terms as being explanatory of observables, presupposes the observable-nonobservable distinction. Secondly, introducing abstract entities and properties is problematic because it is quite unclear what this means. These entities and properties are not abstract in any ordinary sense. Certainly those that are observable, are very much concrete, in the normal sense of that word. And how are the referents of theoretical terms removed from the phenomena? Is the heart, understood as the referent of a term in a biological theory of the human organism, 'removed from, and underlying' the body?

says 'evidential phen.' not 'observables'

One way to try to save Gopnik, Meltzoff, and Wellman's view, is to rephrase the distinction in terms of 'observed' and 'unobserved'. At the time the theory is constructed, theoretical terms refer to properties, entities, etc., that are unobserved at that time, and that are explanatory of a group of observed properties, entities, etc. However, I think this is unsatisfactory for a number of reasons. One is that it is somewhat cumbersome to have to look at the history of a body of information, in order to determine whether or not it is a theory. More importantly, it sits somewhat awkwardly with the structure of science. What counts as theoretical entities at one level of description, can count as forming part of the data at some other, lower, level of description. For example, cell-division plays a causal-explanatory role in biology. We should therefore regard cell-division as a theoretical posit in biology. However, in chemistry, cell-division might play the role of the fact that needs to be explained - the data. If we describe theoretical posits as unobserved entities, properties, and so on, and the data as what is observed, we end up with the rather unnatural idea that what counts as observed at one level of description - in one scientific theory - counts as unobserved at another level of description - in another theory. Therefore, I suggest that we look at it this way.

Doesn't sound unnatural to me.

[31] In Gopnik (1993a), she says that "[c]hildren's theories of mind postulate unobserved entities (beliefs and desires)" (p. 10). However, elsewhere she is careful to talk of 'abstract entities' as opposed to 'unobservable entities' when talking of psychological states (Gopnik & Wellman, 1992 & 1994; Gopnik & Meltzoff, 1997).

Theoretical posits are entities, properties, magnitudes, and the like, that explain some aspect of the world at some level of description, for example by being causally efficacious in the production of the relevant aspect. A prototype of such entities and properties are unobservable entities and properties, but their defining characteristic is their role in explanation. The contrast will be not between observed and unobserved, but between the role that the relevant entities, properties, and so on, play. What counts as theoretical posits depends on what counts as data, and *vice versa.* This way of regarding matters fits the structure of explanation, as we know it, much better. We can go rather a long way explaining one thing, and then explaining the thing that explained the thing, and so on. Some hold that this kind of explanatory reduction may be taken all the way down to basic physics. However that may be, it is easy to see that what counts as the data for one theory might play the explanatory role in another. Gopnik, herself, says as much in an earlier paper of hers: (Gopnik, 1988, p. 198-99)

> Typically, we assume that the evidential level of description is somehow closer to basic perceptual information than the theoretical level. However, contemporary philosophy of science suggests that there is no bottom-line uncontaminated evidential level, no ultimate foundation on which theoretical structures rest. Similarly, there is no hard and fast line between the evidential and theoretical level of description. Often one theoretical description can function as the evidential base for a higher level theory.

A theoretical posit, then, is a functional construction in the sense that good it is only a posit relative to a particular framework of explanation. Whereas it is not is not clear that all scientific explanation is *causal* explanation, as (ii) says, it is certainly true that many, perhaps most, scientific theories contain theoretical posits that explain the data causally. Therefore, it does not seem unreasonable to refer to causal explanation here. It should be stressed, however, that causal explanation is just a subclass of explanation more widely.

(i) might seem unproblematic to start with, but actually betrays a strong commitment to a semantic assumption: that terms can be defined. There is a particular theory of the meaning of theoretical terms, that I suspect is what lies behind (i): the Ramsey-Carnap-Lewis theory (Carnap; 1956; Ramsey, 1978; Lewis, 1972). The basic idea is that theoretical terms can be non-circularly defined in terms of the statements of the theory in which they figure. In a theory, one might individuate two kinds of terms; theoretical terms and other terms that form part of the formulation of the theory. The latter might, for example, be terms that refer to the data. Let us call these T-terms and O-terms respectively. In a sentence that states the theory, you replace all T-terms by unbound variables. You then bind them with an existential quantifier, thereby specifying that the theory has at least one realisation.[32] Through further operations, we eventually end up with a sentence in which the T-terms are defined by the relation that they bear to each other and to O-terms. In this way, the meaning of T-terms depends only on their interrelation and the meaning of the O-terms. Once one knows the meaning of the O-terms, the meaning of the T-terms can be extracted from how these terms figure in generalisations or laws involving them. This implies that theoretical terms are interdefined. They are defined according to the relation they bear to each other and the O-terms.

However, we might ask ourselves whether we should accept this theory. There are other possible views about the meaning of theoretical terms. Most importantly, there is the Kripke-Putnam view of the meaning of physical magnitude and natural kind terms (Kripke, 1980; Putnam, 1973 & 1975a). According to it, the meaning of physical magnitude and natural kind terms is primarily given by their reference. This view might be extended to cover other theoretical terms also.

Inspired by Saul Kripke's theory of proper names (Kripke, 1980), Hilary Putnam has suggested that physical magnitude terms get their meaning in the following way (Putnam, 1973). Some kind of introducing event or baptism takes place, in which a physicist picks out a magnitude by a definite description. This description will normally, though not necessarily, be a causal description - the magnitude is picked out by reference to its observable causes or effects. Once the person propagates the term, all uses of it will be connected by a causal chain leading back to this initial event. The

Priority;

[32] Lewis (1972) goes on to bind the variables uniquely. I will not go into detail with these technicalities here.

meaning of the term, then, is not given by a set of necessary and sufficient conditions that a speaker must know. The only thing Putnam can think of that all speakers need to know, is that a physical magnitude term refers to a physical magnitude, something that allows of more or less and that is capable of location (p. 199). This information may form part of the meaning of the term, but it is chiefly the reference of the term that fixes its meaning. Thus, rather than speakers having to *know* certain things, they need to stand in certain distinguished relations to the referent of the term - they need to be hooked up to the right causal chain, as it were. All kinds of descriptions may be linguistically associated with the term, but they do not provide its meaning, if by that one has in mind necessary and sufficient conditions that speakers must know. Primarily what all uses of a physical magnitude term have in common, ~~are~~ is that they all refer to the same magnitude.

Now, we can imagine a similar account to that of the meaning of physical magnitude terms applying to other theoretical terms. As such, it should be clear that it isn't necessary to commit oneself to a strong view of the definability of theoretical terms. The Kripke-Putnam theory of the meaning of theoretical terms has the advantage of requiring less knowledge on the part of speakers, which makes it more plausible. However, I will not here adjudicate between the two. I do think, however, that caution is required at this point, and that it is incautious to commit oneself to a strong account of the definability of theoretical terms. Instead I suggest a weaker expression of what I take (i)? to be the basic point behind (ii). Theoretical terms do seem to be interlinked in some interesting fashion. For example, our understanding of 'quark' is related to an understanding of 'leptons', 'protons', 'photons', and so on. An understanding of 'photons', in its turn, is related to an understanding of 'quark', 'leptons', 'protons', and so on. The more you know about the meaning of a theoretical term, the more you are likely to know about the meaning of other theoretical terms of that theory. Another way of putting the same point, is to say that theoretical terms form a coherent structure, where one theoretical term is connected to other theoretical terms of the same theory in such a fashion that normally, understanding it involves understanding some or all of the others. This is a weaker claim that to say that they are interdefined. Prudence should lead us to embrace this option.

(iii) may also be in need of some amendment. Some philosophers mean 'strict exceptionless laws' by 'laws'; Davidson, for example (Davidson, 1974). By this standard, at most physics will count as a scientific theory. Whereas it is certainly true that physics is commonly regarded as the prototypical science, containing the strictest laws around, it is not true to say that it is the only scientific theory. Laws in biology, geology, astronomy, and so on, are not exceptionless, nevertheless they are generally agreed to be scientific theories (cf. Fodor, 1987). However, some, like Cartwright, argue that there are no "exceptionless quantitative laws in physics" (1983, p. 46) either - quantitative laws being prototypical physical laws. According to her, not even the law of universal gravitation is exceptionless. The problem is that many forces, other than that of gravity, are at play at any time. The law of universal gravitation only describes how mass is affected by gravity, if there are no other forces acting on that mass. When we calculate the force acting on charged particles, for example, we must factor Coulomb's law into the equation. The law of universal gravitation is an idealisation that is actually never true of any natural phenomenon. According to Cartwright, it is only true, if it is expressed in these terms: (1983, p. 58)

> *If* there are no forces other than gravitational forces at work, *then* two bodies exert a force between each other which varies inversely as the square of the distance between them, and varies directly as the product of their masses.

This is no longer an exceptionless law because the antecedent is a *ceteris paribus* clause. It specifies the conditions under which the consequent is true, namely when there are no other forces than the gravitational one at work. And, since there are many other forces than gravitation at work at any time, it is unlikely that there are any circumstances where the law is applicable. Because the antecedent never holds true in nature, Cartwright regards the law as never actually being true of anything (other than counterfactuals).

counter-factual situations.

Someone might argue that the law of universal gravitation is always true, since every mass is affected by it according to the parameters set out in that law. The fact that there are also other forces at work and, consequently, the resultant force is different from that predicted by the law of universal gravitation, is neither here nor there. However, Cartwright's point is that this way of regarding laws makes them connected with causal *powers* rather than with actual events. This is unfortunate because "[w]e need an account of what laws are, an account that connects them, on the one hand, with standard scientific methods for confirming laws, and on the other, with the use they are put to for prediction, construction, and explanation" (1983, pp. 61-62), and there is no way of providing such an account in terms of either causal powers or laws. Cartwright, herself, favours viewing nature in terms of capacities, but this she takes to exclude talk of laws (Cartwright, 1989).

If one wishes to talk about laws, one must accept that, if they are taken to describe facts, they are almost always nowhere close to being exceptionless. Perhaps there are some non-quantitative laws of physics that are exceptionless. The fact remains that the vast majority of the laws of physics are quantitative. However, as we have seen, either quantitative laws describe facts and are not exceptionless, or they are exceptionless but don't describe facts. I take this as being quite a strong argument against regarding the laws of physics, in general, asbeing exceptionless.

To cut a long story short, there are reasons not to include strict, exceptionless laws among the characteristics of scientific theories. Firstly, not even physics trades much in such laws and, secondly, other sciences, such as biology and geography, do not contain such laws. Therefore, I suggest that we reformulate (iii) such that it does not
? talk of laws, but of lawlike relations. There are at least two ways in which one might regard lawlike relations: in terms of causality, or in terms of counterfactuality. Everybody agrees that there is quite a close relation between the two. However, some believe that causality can be analysed in terms of counterfactuals (Hume, 1777/1975; Lewis, 1973? Fodor, 1987). 'If there *were* a black hole in our solar system, we would be pulled into it' is a counterfactual supported by the law of universal gravitation and probably a handful of other laws too. Others take the concept of causality to be the basic one in explanations of natural phenomena (Cartwright, 1989). In this case, causal relations form the basis of all scientific theorising. If, however, we take seriously the suggestion that not all explanations are causal explanations, that might give us reason

to think that some laws are better described in terms of counterfactuals than in terms of causality.

(iv). The idea that scientific explanation exhibits a deductive-nomological (D-N) structure amounts to the following idea. We explain a particular event by specifying certain initial conditions, and by pointing to a law that links these initial conditions with the event to be explained. For example, this ball accelerates at *x* mph [the explanandum], because its mass is *y*, and it undergoes the force *z* [the initial conditions], and F = ma [the covering law]. The combination of these statements is what constitutes a D-N explanation. The explanatoriness of the explanation lies in the fact that the particular event is shown to fall under a law (cf. Hempel, 1965, Essay 12).

Most theory theorists explicitly adhere to the D-N picture of explanation - for example, Churchland, Fodor, Gopnik, Meltzoff, and Wellman - and I know of no theory theorist that has advocated another account of scientific and folk psychological explanation. Most simulationists also believe that theory theorists hold that folk psychological explanations are D-N explanations, and have been concerned to provide an alternative account of explanation since simulationism does not lend itself to such explanations (Gordon, 1992a; Heal, 1998, Davies & Stone, 1998).

However, the D-N picture of explanation is not unproblematic. Michael Friedman (1974), for example, complains that it isn't really clear what is explanatory about D-N explanations. The best Hempel does for the idea is to say that, given the initial conditions and the law, we would expect the relevant event or state of affairs (Hempel, 1965, p. 327). However, there are many cases in which one might expect a state of affairs based on knowledge of some law and initial conditions, without thereby having gained an understanding of it. For example, I may be able to predict the storm from reading the barometer - the law I use in this instance is known as an indicator law. Clearly, subsuming the event of the storm under an indicator law, does not enhance my understanding, and hence does not provide an *explanation*. Furthermore, D-N explanations fall short when it comes to explaining general regularities. Since such regularities do not occur at definite times, their occurrence cannot really be expected (Friedman, 1974, pp. 8-9). These and other objections (see Lipton, 1991) indicate that, at best, the D-N model provides necessary but not sufficient conditions of scientific explanation.

On the other hand, there seems to be no other entirely satisfactory picture of scientific explanation. Peter Lipton (1991) has suggested that we explain *x* by pointing to what caused *x*. However, as it is unclear that *all* scientific explanation is causal explanation, he doesn't consider his account to be a complete one.

If we reconsider (ii) at this point, we see that Gopnik, Meltzoff, and Wellman actually talk about causal explanation also. What they might have in mind is that the covering laws on the D-N model are causal laws. We can then regard Gopnik, Meltzoff, and Wellman as endorsing a view where explanation is both deductive-nomological and causal. Should we accept such a view? Now, as we have seen, it is not immediately obvious that all scientific explanation need be causal explanation. A more careful formulation of the basic idea might go like this. A theory provides explanations phrased in terms of certain entities and properties that either cause the data and/or are correlated to them in a lawlike manner. This may still seem a bit strong, but if we embrace anything much weaker, we could end up with too loose a notion of theoreticity.

constitution?

indicator laws eg.

Now that we have examined all of (i)-(vii) in more detail, and have excluded and amended conditions where deemed necessary, we can now reformulate the conditions of theoreticity. The conditions, as I will understand them, are jointly sufficient and individually close to being necessary, *ceteris paribus*, for some body of information being a theory.

Nat selection?

> A body of information is a theory if:
>
> (a) it consists of a number of lawlike generalisations, and
>
> (b) these generalisations contain terms that refer to entities and properties that explain some data, for example by being causally efficacious in the production of them, or by being related to them in some lawlike fashion, and
>
> (c) the terms in (b) form a coherent, interrelated structure.

explicate

explicate?

Are instrumental theories, theories. Seems to be committed to realism.

(a) and (b) contain elements of (ii), (iii) and (iv), and (c) is a reformulation of what I have taken to be the basic idea behind (i). (v)-(vii) were discarded straight of. This picture presupposes some understanding of explanation that is independent of theory. However, I don't regard this as being a serious difficulty. I take it that this is an account that philosophers of science are able to provide. There is, at any rate, no reason to believe that the project is impossible. The question now is whether these features fit with those exhibited by the body of folk psychological knowledge. That is the topic of the next section.

2. The Theoreticity of the Body of Folk Psychological Information

The idea of the continuity between common sense and science is rather appealing, but does it extend to (a) - (c) holding true of the body of folk psychological generalisations?

What sense of FP [illegible] [illegible] is used here? Int or Ext? If Int, how do we know what they are?

(a) Does folk psychological theory consist of lawlike generalisations? No theory theorist wants to hold that folk psychological information is contained in mere generalisations. This would fail to distinguish uninteresting and accidental, but true, generalisations like 'if *x* is in this room and wearing a blue T-shirt, then *x* is a female' from interesting, non-accidental generalisations like 'if *x* is an atom, then *x* has a nucleus'. This distinction can be kept in place by distinguishing between generalisations and counterfactual-supporting generalisations. The former can express any kind of correlation - however trivial and accidental, like the correlation between being female and wearing a blue T-shirt. The latter expresses only non-accidental generalisations, like that concerning the atom. Non-accidental generalisations are likely to express something about regularities in nature, whereas accidental ones are not. Another way to capture part of the interesting fabric of nature is in terms of causal generalisations - generalisations that say something about how things come to be the way they are. As we saw above, either of these ways of regarding the relations that folk psychological generalisations map, can be described as being lawlike. Hence, (i) seems to fit nicely with folk psychology.

Nevertheless, philosophers like Donald Davidson (1974) and Kathleen Wilkes (1984) have argued that there is a big difference between the *ceteris paribus* clauses of physical theory and those of the body of folk psychological information. Those of physics can be filled out, at least in principle, whereas those of folk psychology cannot. According to Wilkes, if you try to fill our the *ceteris paribus* clauses of the generalisations of folk psychological theory, you end up making them true of almost anything. It is not quite clear why she thinks so. Presumably, she cannot mean that simply because the antecedent fails to be satisfied in a great number of cases, the generalisations are (counterfactually) true of all of those cases, for the same holds true for the laws of physics. If she means that the generalisations are applicable to almost anything, this is blatantly false, since there are countless events and states of affairs that the generalisations of folk psychology do not apply to. One problem might be that she concentrates almost exclusively on proverbs like: 'too many cooks spoil the broth', 'out of sight, out of mind', and 'absence makes the heart grow fonder' (pp. 344-46). As she points out, some of these proverbs are mutually contradictory as they stand, and one can imagine that once the relevant *ceteris paribus* clauses are spelt out, the proverbs will have become vacuous (but see Furnham, 1987).

Int or ext?

However, I don't think that proverbs are properly seen as forming a part of the core of the body of folk psychological information. Some proverbs don't even contain psychological terms. All of these we can exclude from the body of folk psychological information straight away. There are good reasons to exclude those that do contain such terms as well. Firstly, in general, proverbs are taken with a pinch of salt - few people believe that they are actually true. They are more like slogans that can profitably be used under certain circumstances. Folk psychological generalisations are taken at face value. ? Secondly, proverbs generally need interpretation as they are not transparent. They are most commonly heavily metaphorical. When you stay out of a decision making process, justifying your action by the proverb 'too many cooks spoil the broth', you are not concerned with either cooks or broth. On the other hand, when you quote a folk psychological generalisation, like the action generalisation, you are precisely occupied with belief, desire, and action. Thirdly, folk psychological generalisations are properly explanatory, proverbs aren't. This might have to do with the difference in rationality of the two. Folk psychological generalisations make sense in a way that proverbs don't. Proverbs just plot certain correlations that are sometimes observed to obtain. The facts that many cooks spoil the broth or that absence

?

makes the heart grow fonder are not explanatory in the way the action generalisation is. Fourthly, proverbs often have opposites. For example, 'absence makes the heart grow fonder' is the companion to 'out of sight, out of mind'. I cannot think of any folk psychological generalisations, when their *ceteris paribus* clauses are specified, that contradict each other in this fashion. Fifthly, proverbs only make sense against the background of a lot of other information that we have about people. Take 'out of sight, out of mind'. The fact that one might be less occupied with something to which one's attention isn't constantly drawn by its presence, is information that folk psychological theory proper will give us. It is worth looking into the idea that proverbs only make sense against the background of folk psychological generalisations. Unfortunately, I cannot do such a project justice here. However, given this list of dissimilarities between proverbs and folk psychological generalisations as described in chapter 1, I think it fair to conclude that proverbs do not form part of the body of folk psychological information proper.

See Elster on proverbs that contradict it.

This is not to say that proverbs play no role in our psychological attributions. They might, but that doesn't make them form part of the body of folk psychological information proper. Much information other than that contained in this body, plays a role in our psychological attributions; for example knowledge of folk physics, public affairs, and so on. This information does not form part of the body of folk psychological information either.

If we think back at the examples of folk psychological generalisations in chapter 1, we see that they do have the form of laws of the special sciences. They are counterfactual supporting generalisations. Like special science laws, they are not exceptionless. However, this does not mean that they are vacuous, since we have no reason to think that their *ceteris paribus* clauses cannot be spelt out in principle (ditto for special science laws). In chapter 1, we saw the beginnings of such a spelling out. Therefore, I think we can safely conclude that the body of folk psychological information consists of a number of lawlike generalisations.

What supports the claim that these are FP gens?

(b) Do folk psychological generalisations contain terms that refer to entities and properties that explain some data, for example by being causally efficacious in the production of them, or by being related to them in a lawlike fashion? If we take a prototypical folk psychological generalisation, like the action generalisation, we find such terms as 'belief', 'desire', and 'does something' (a rewriting of '*A*s'). The ~~latter~~ last is a description of the datum special to the theory. There are other ways in which the datum could be described; as a movement, for example. 'Belief' and 'desire' refer to psychological states that are causally explanatory of the datum under the particular description. Beliefs and desires are related to the data - action - in a lawlike fashion. It is easy to understand folk psychological explanation in a D-N fashion. Why does John move the garden chairs inside? Because he believes that it will rain, that in order that people have something to sit on, he needs to move the chairs inside, and he wants people to have something to sit on. Generally, when people want other people to have something to sit on, and they believe that it will rain, and they believe that if they move the garden chairs inside, people will have something to sit on, then, *ceteris paribus*, they will move the garden chairs inside. Generally, we don't actually go through such a cumbersome explanation. We have explanations like, 'because he wanted people to have something to sit on'. This is a simpler causal explanation. However, we can understand such explanations as short-hand for the more cumbersome D-N type of explanation. In short, folk psychological generalisations contain terms that refer to entities that are causally efficacious in the production of behaviour, and are also related to behaviour in a lawlike fashion. Therefore, we are justified in concluding that (b) holds true of the body of folk psychological information.

However, it has been argued that there are important differences between the referents of theoretical terms of the sciences and the referents of folk psychological terms. One such criticism comes from Adam Morton, who argues that the difference is such, that we are not licensed to call the body of folk psychological information a theory (Morton, 1980).

According to Morton, theoretical terms in the sciences have determined referents. An atom simply is an atom - there is only one natural kind that this term refers to. As opposed to this, "schematic terms" - schemes being higher-order implicit bodies of information - do not have determinate referents. Folk theories are schemes rather than theories. We can see this because: (pp. 28-9)

> people with radically different conceptions of the mental, dualists and materialists, bishops and their neurologists, can easily recognize the shared allegiance to a common-sense conception of the mental, that allows them to discuss motives and characters.

In other words, although there is agreement about how the mind *works*, there is little agreement about what the mind *is*. This is not dissimilar to Atran's view of biological categories (chapter 2, section 1). Put differently, folk psychological terms allow for several types of naturally occurring states or properties as their referents. Other fields of study are also characterised by this feature, for example phonetics and linguistics. (The word 'dog', is sometimes spoken, sometimes written, and may be writen and spoken ways that are distinguishable at the level of phonetics or graphics.) Like the body of folk psychological information, they are best regarded as schemes.

The different ways in which schematic terms and theoretical terms refer, have consequences for how the bodies of information respond to new evidence. Imagine that you are a type-type identity theorist. You believe that pain is identical to C-fibres firing. Then you read Putnam's "Philosophy and Our Mental Life" (1975a). He argues that pain is really a functional state. He asks you to imagine meeting Martians, and discovering that they are functionally identical to us in all respects, but have a different brain chemistry, and hence no C-fibres. Would you deny that they were in pain under the same kinds of circumstances where we are in pain, given the functional isomorphism between us? Putnam claims that this would be absurd (p. 293). You agree. What appears to be the essence of pain is not the particular realiser of the pain - C-fibres firing, or what have you - but the complex of its typical causes and effects. It is caused by the body being bruised, compressed, burnt, pierced, etc., causes a characteristic experience, sometimes wincing, crying, and often some kind of evasive behaviour. This realisation causes the denotation of the term, rather than its meaning, to change. That is, we extend the term to cover Martian pain, rather than deciding to deploy it exclusively to refer to the cases that we have always thought were cases of pain.

Morton claims that a similar kind of situation would provoke quite a different outcome in any scientific discipline. If it were discovered that, for example, chemical compounds were not composed of molecules, but of something else, and that all that chemistry says is true of molecules is true of these other things, we would not dub these other things 'molecules'. Molecules "would still be the little congeries of atoms they have always been" (pp. 22-3). In other words, in science we change our concepts when confronted with new evidence of this sort, rather than extending the relevant concepts to the newly discovered evidence.

It would appear, then, that a term that has a 'determinate referent' is really a rigid designator. Indeed, it is hard to see what else Morton could have in mind. Consequently, we can reformulate Morton's idea thus: bodies of information whose terms non-rigidly refer are not theories because terms of theories are rigid designators. However, if this is what Morton has in mind, his conclusion doesn't follow. The point is that rigid designators only designate determinate kinds at their own level of description. There will always be some level of description at which they do not rigidly designate anything unless, that is, one believes in an ultimate, irreducible level of description.

Let us take Morton's own example of molecules. At its own level of description, the term 'molecule' rigidly designates just one kind: molecules. However, at any lower level of description, 'molecule' does not rigidly designate anything. Take the level of physical theory. It is quite possible that in some possible world, "dwarks" and not quarks form part of the composition of molecules. At this level, then, 'molecule' does not rigidly designate. Rigid designators only rigidly designate objects, events, or states of affairs, at some particular level of description. Following what Kripke has said about psychological terms, they are themselves rigid designators. 'Pain' designates pain in all possible worlds (Kripke, 1980). However, at the level of neurophysiology, 'pain' does not rigidly designate anything. This parallels the case with 'molecules'.[33]

cf Twin Earth intuitions.

[33] The molecule example is puzzling in a number of ways. Why does 'molecule' continue to denote little congeries of atoms? That is, isn't the fact that 'molecule' denotes little congeries of atoms one of the claims of chemistry that we supposedly discover is true of something different from what we thought it was true of (and that we have called 'molecules' so far)? And if it is, how can we decide that 'molecule' continues to denote little congeries of atoms after the discovery that Morton envisages? How, in any case, is it discovered that molecules exist? Compare with 'atoms'. It was discovered that matter is not made up of indivisible atoms. Yet we retained the term 'atom' and revised our concept. In order to make his example carry the weight he wants it to, Morton needs to answer all these questions satisfactorily.

Morton's criticism, then, does not really amount to a defeat of this line of Theory Theory. The most it does, is the following. Assuming that there is an ultimate, irreducible level of description, and assuming that it is that of physics, the terms of theories in physics always rigidly designate kinds. There will be no other level of description at which they do not rigidly designate something. However, Theory Theory is not interested in a parallel with physics as such, but with scientific theories more generally. Scientific theories include sciences like biology, geology, anatomy, and so on. And the terms of these theories are such that, although they may rigidly designate something at one level of description, there will always be another at which they don't. Just as in the case of folk psychological terms. To conclude, Morton's objection does not constitute a real threat to the line of Theory Theory that we are concerned with here. Therefore, there are very good reasons to think that folk psychological terms are theoretical terms of the type described in (b).

H_2O"

(c) Do psychological terms form a coherent, interrelated structure? They appear to. A belief is a psychological state that purports to covary with the environment, and that, together with desire, cause action. Action, on the other hand, is a kind of behaviour that is caused by a belief and a desire, or some other combination of psychological states (for example a fear and a belief). An emotion, like grief, might be caused by the belief that someone loved has died. This belief, in its turn, might be caused by certain perceptions - for example seeing the beloved pale and still at the morgue. And so on. Understanding one psychological term seems to require understanding at least something about other psychological terms. And learning more about one psychological state, feeds into what one knows about other psychological states. Just like theoretical terms in the sciences, psychological terms seem to form a coherent, interrelated whole.[34]

Another objection one might raise, is that Morton seems to take it for granted that a Kripke-Putnam theory of the meaning of theoretical terms is true. However, if one accepts the Ramsey-Lewis theory, then, after the discovery that Morton envisages, either 'molecule' comes to designate the newly discovered things or, alternatively, we would say that there are no such things as 'molecules'.

[34] Jan Smedslund (1990) has argued that since psychological terms are interdefined, all the laws of folk psychology are *a priori*. If you exchange a term for its definition in the formulation of a law, it becomes tautological. This is not a million miles away from Lewis' suggestion about how to define psychological terms (Lewis, 1972). For more on Lewis, see chapter 4. I will discuss Smedslund's work in more detail in the conclusion.

We can now conclude that the body of folk psychological information is a theory, since it possesses three characteristic features of scientific theories that are jointly sufficient and individually close to being necessary for something being a theory. The body of folk psychological information consists of a number of lawlike generalisations that contain terms that refer to entities and properties that explain some data, either by being causally efficacious in the production of them, and/or by being related to them in a lawlike fashion. Psychological terms also form a coherent, interrelated structure.

However, before settling on this version of Theory Theory, we need to look at another influential idea of theoreticity that folk psychological theory can be modelled on: the Kuhnian model.

3. Folk Psychological Theory as a Framework Theory

The second view of how the body of folk psychological information compares to scientific theories, is also propounded by developmental psychologists, such as Wellman (1990) and Carey (1985). According to them, theoreticity is best modelled on a largely Kuhnian idea of science - in terms of paradigms or, as Wellman calls them, framework theories. More importantly, the development of folk psychological knowledge is modelled on Kuhn's idea of scientific revolutions, or paradigm shifts (Kuhn, 1970a, 1970b).[35] It is possible to regard this model of theoreticity as the developmental counterpart to the more static model just discussed. It seems to be the view that Gopnik & Meltzoff, and Wellman, in his later work co-written with Gopnik, take. I think there is a real question as to the extent to which a Kuhnian picture of science can be reconciled with a more traditional account, even considering that Kuhn concentrates on the dynamic features of science. I cannot, however, examine this idea in detail here. Instead, I shall simply regard the framework theory as providing an entirely separate model of theoreticity. This is justified by both Carey (1985) and Wellman (1990) understanding the theoreticity of folk

[35] The philosophies of I. Lakatos and L. Laudan also enter into the picture, but since I cannot do justice to all these influences, I will concentrate on the Kuhnian one.

psychological knowledge exclusively in such terms. Indeed, Wellman (1990) rejects any other way of drawing the parallel between scientific theories and everyday theories (pp. 123-5). No doubt there are some differences between Wellman's and Carey's views of theoreticity, but I shall concentrate on what I take to be the commonalities.

In *The Child's Theory of Mind*, Wellman says that framework theories "define the ontology and the basic causal devices for their specific theories and even constrain some aspects of accepted methodology" (p. 125). Framework theories are more global in their scope than specific theories; they are sufficiently underspecified to stimulate research, whilst still being powerful enough to sustain an entire research tradition; they direct the theorist's attention towards certain kinds of phenomena rather than others; they indicate what kinds of questions it is legitimate to ask, and what counts as answers to them.[36] Examples of paradigms are Ptolemaic astronomy, Copernican astronomy, Newtonian mechanics, and quantum mechanics. There was a change of paradigms between the first and the second, and the third and the fourth. Within each of these paradigms normal science takes place. Specific theories are developed inside the paradigm. For example, quantum field theory and quantum electrodynamics are specific theories of the quantum mechanics paradigm.

Wellman focuses primarily on 3 features of framework theories: their ontology, causal features, and methodology. These are features of the body of information at the core of a paradigm that regulate how future theories can develop. When Wellman says that a framework theory dictates *ontology*, I take it that he means that it states what entities and properties form part of a certain domain, either directly, or by being causally active in the production of events and states within it.[37] A theory, then, must first delimit what counts as data, and by reference to these, point to what counts as elementary entities and

[36] In this formulation, Wellman seems to be much closer to the view of paradigms that Kuhn defended in *The Structure of Scientific Revolutions*, than the one he subsequently espoused after it was pointed out to him that the notion of 'paradigms' was impossibly vague. According Margeret Masterman (1970), Kuhn (1970a) used the term in at least 21 different senses. In his (1969) postscript, he coined a new phrase 'disciplinary matrix' that covers the part of the sense of 'paradigm' that referred to the idea of science as puzzle solving within a community in his (1970a). However, this seems somewhat narrower than what Wellman has in mind.

[37] It cannot have anything to do with what really exists, for it has often been argued that Kuhn is an anti-realist. Had he propounded clearly realist views about the objects of science, somebody would surely have noticed. ?

K. claimed to be a realist.

properties. In quantum mechanics, for example, the elementary entities are subatomic particles, and elementary properties, properties of subatomic particles. The theory also determines the way in which these elementary entities or properties can interact - for example, by allowing or disallowing action at a distance. This is what is referred to by *causal features*. Lastly, there is the *methodology* of framework theories. Here we are presented with the view that whereas changes in specific theories are "data-driven and can be regarded as progressive" (p. 126), changes in framework theories cannot. ? Specific theory change is governed by the basic commitments of the framework theory. One way in which theories are commonly evaluated is by reference to their puzzle solving abilities. It is relatively straightforward to evaluate the explanatoriness of specific theories. But Kuhn (1970a) claims that framework theories cannot be evaluated in this fashion. One cannot choose one paradigm over another in accordance with its puzzle solving abilities, for no paradigm ever solves all puzzles that are accepted as reasonable within it, nor are the puzzle solving abilities of different paradigms comparable since they often solve *different* puzzles. What counts as a puzzle for one paradigm, might be a non-puzzle for another. Moving to a Newtonian picture, gravity came to be seen as something basic that was not to be explained by something else in its turn. Earlier, however, it was regarded as something to be explained, and a magnetic theory was proposed for this purpose. So, with shifts in paradigms, puzzles appear and disappear. The particular example of the paradigm shift to Newtonian theory, is a case where, in one important respect, the new paradigm does not have more puzzle-solving power than the old one.

Kuhn, however, is not always as pessimistic about the rationality of paradigm change, as Wellman might lead us to believe. There appear to be five important features that are generally - that is, cross-paradigmatically - regarded as being features a good theory should possess: accuracy, consistency, broad scope, simplicity, and fruitfulness (cf. Newton-Smith, 1981, p. 113). They play a role both in the choosing and the formulation of paradigms; they constrain what the scientific community will regard as a suitable paradigm. According to Kuhn (1970a), when there is a paradigm shift, scientists belonging to the old paradigm and scientists adhering to the new one, cannot communicate because the basic concepts have changed. One of the basic functions of a paradigm is to allow discussion among scientists. This is only possible against a background of agreement as to what kind of things and events count as problems for science, what counts as solutions, what are taken to be basic, irreducible phenomena, and so on. Paradigm shifts involve incommensurable world views (Kuhn, 1970a, p. 150) - completely different concepts. Paradigm shifts are provoked by a crisis in a paradigm. Too much counter-evidence, too many *ad hoc* hypotheses, and so on.

The idea of paradigm *shifts* is probably what has attracted most developmental psychologists to the framework theory idea. Carey is primarily concerned with describing the ontogenesis of our understanding of biological and psychological categories. She compares an earlier stage of such understanding - the one that young children have - with a later one, the one that adults possess, and concludes that there are good reasons to think that there has been a conceptual change. For Carey, the conceptual change that she believes occurs in childhood is best modelled on paradigm shifts. Changes in mere (specific) theory can often be restricted to some conceptual *enrichment*. The conceptual change that she has researched seems much more radical. It is not just a matter of concepts becoming richer, as frequently occurs in learning, but that concepts *change*. This is best modelled not on simple theory change, but on paradigm shifts.

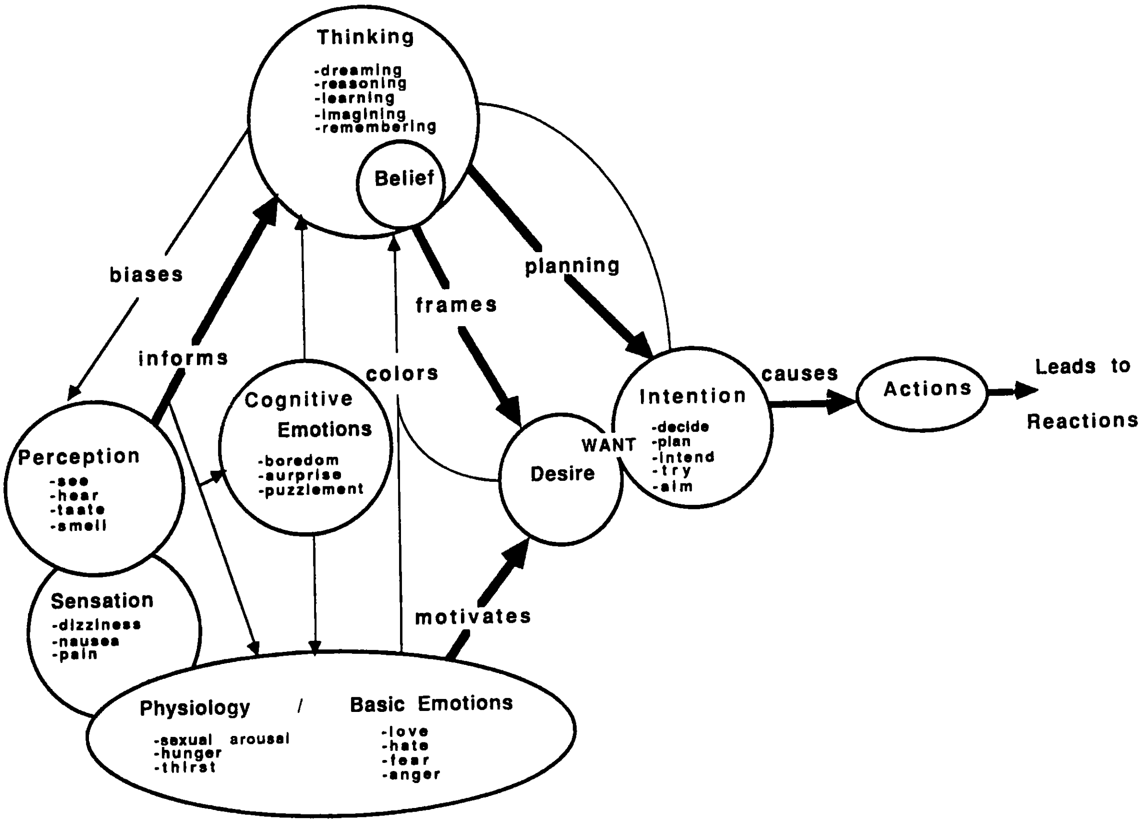
Both Carey and Wellman reject Kuhn's early formulation of paradigm shifts in terns if incommensurable world views. Carey prefers to think in terms of 'local incommensurability', a later Kuhnian approach. According to this: (Kuhn, 1982, p. 670)

> The claim that two theories are incommensurable is then the claim that there is no language, neutral or otherwise, into which both theories, conceived as sets of sentences, can be translated without residue or loss.

This is considerably more modest than a 'different world views' approach. It seems that quite a bit of communication will be possible. The approach that Wellman adopts is similar to this. There can be *some* communication between proponents of different paradigms. Although the Representational Theory of Belief is very different from the Copy Theory of Belief, for Wellman it is clear that elements remain.

For example, beliefs are internal states of people related to motivational states and reality in some fashion.

Wellman takes, as far as I can see, much of what I have called the core of folk psychological theory and construes it as a framework theory. His figure 4.2 (p. 109) provides some idea of what he has in mind:



I will not go into further detail with Wellman's exposition of (some of) the core of folk psychology. I don't believe that it is necessary, because there are already problems at the level of characterising scientific theories. If the framework theory does not fit well with scientific theories, the fact that folk psychological theory could be made to fit with this model wouldn't show what it was a theory.

Kuhn's view of science is disputed. The idea of a paradigm, for example, is vague, even when reformulated. It is extremely difficult to pinpoint paradigms in science (Newton-Smith, 1981). This is no coincidence - it is just very hard to know exactly what to count as a paradigm. I, however, do not intend to criticise the framework theory model on this point. Instead, I want to focus on the nature of scientific theory *qua* theory. The fact that this issue is never properly addressed is, to my mind, the most serious shortcoming of this model of theoreticity. The notion of theory seems to be taken for granted, and the model is one of particular *kinds* of theories. There are two kinds of theories: specific theories, whose theoreticity is not discussed, and paradigms, which are quite different from what we would normally think of as theories; they are research traditions or shared assumptions about how to solve the puzzles of nature. Wellman recognises this, but claims that: (p. 127)

> a particularly useful level of analysis for finding similarities between scientific and commonsense theories is that of framework theories. A level of analysis more replete with dissimilarities is the comparison of commonsense and scientific specific theories.

This model of theoreticity seems to be much more specific than that discussed above. According to the traditional model, the body of folk psychological information is a theory because it shares certain interesting features with scientific theories in general. According to the framework theory, the body of folk psychological information is a theory because it is a theory of a special kind: a framework theory. This is a much stronger claim. Since all theories have to have certain characteristics in common, particular kinds of theories must have characteristics over and above those that make them part of the class of theories. There are two problems here: how could holding this view possibly be a weaker view of the theoreticity of the body of folk psychological information than a traditional one? And why would theory theorists want to hold anything so strong?

What makes specific theories theories and what makes specific theories and framework theories both theories, are never explicitly discussed. Therefore, I feel justified in construing the theoreticity of both on the model of scientific theory that we have discussed above. If we assume this, then the framework theory comes out as a much stronger view than the traditional model of theoreticity. This does make Wellman's claim that the framework theory is a weaker view than a traditional picture rather mysterious. For if specific theories and framework theories are both scientific theories in the above sense, and folk psychological theory is a framework theory, then folk

The trad view is an account of specific theories. The conditions on being a specific theory are stronger ~~weaker~~ than those on being a f'work theory.

psychological theory *will* have interesting features in common with specific theories. This, of course, is a reason for a theory theorist to reject this model. At this point, theory theorists should be interested simply in substantiating the sense in which our folk psychological knowledge is knowledge of a theory, *not* in specifying just what kind of theory it is. This is, no doubt, a worthy enterprise, but something that can be put off until the foundations of Theory Theory are laid down. These problems may seem just to show that my interpretation of 'theory' cannot be what Wellman has in mind. However, since Wellman has not given us any other notion to work with, I cannot see how else to proceed.

It may be objected ~~to me~~, at this point, that I have been unfair to Wellman. If we take the heritage from Kuhn seriously, we can model framework theories neither on folk bodies of knowledge nor scientific bodies of knowledge. A paradigm consists in a number of shared models, shared ideas of what counts as problems and solutions, and so on. Most of these are regarded by Kuhn as constituting tacit assumptions mainly directed at the *application* of science to the phenomena. This idea is certainly very interesting, but suffers from a number of shortcomings. Firstly, it cannot explain why specific theories and framework theories are both *theories*. Secondly, if we admit that calling framework theories 'theories' is a bit of a misnomer, it is infelicitous to use this sense of 'theory' to substantiate the Theory Theory position. Instead, we can imagine it constituting another theory of folk psychology altogether. Note, in this context, that no philosopher of science calls these bodies of information theories: they are called paradigms, research programs or research traditions (cf. Wellman, 1980, p. 125).[38]

good

Although I don't think that the framework theory is a good view of the nature of scientific theories, it is no doubt quite pertinent when looking at science more broadly. It highlights what shared assumptions lie behind the way science is practiced. But however important such features are, they are not features that we need to

[38] The term 'framework theory' was coined by Wellman, himself, to capture subject matters, such as behaviourism, psycho-analytic theory, and so on, not the different specific theories within them as, for example, Skinnerian conditioned response theory and Kleinian theory of the depressive position. This is an interesting project, but it still fails, I think, to show what framework theories and specific theories have in common that make them both theories. And, to reiterate a point made frequently above, it is this common notion of theoreticity that Theory Theory should be concerned with at this stage.

build into the nature of scientific theories *qua* theories. And this is really all the theory theorist should be concerned with at this stage. All that needs to be defended is that the body of folk psychological information is a theory, not whatever other features it may possess in addition, making it more like some form of theory rather than another. Therefore, I propose to reject the framework view of theoreticity.

4. *Other Aspects of Folk Psychological Theory*

Although we should reject Wellman's theory as a model of theoreticity for folk psychology, it is worth pointing out that folk psychology has at least one interesting feature in common with framework theories understood as paradigms. It does not appear to be a theory that one accepts or discards ~~all~~ according to how well it fits the evidence. It seems much more like a world view. In the general run of things, it is not really tested or questioned. It frames the way we conceive of ourselves and others. We don't question the core of the theory if explanations or predictions go wrong, but we may correct generalisations - for example by adding more *ceteris paribus* clauses, or add new hypotheses. We have great difficulties accepting a view of ourselves that excludes us having psychological states roughly as we conceive of them now. The heated debate about whether folk psychology can form the basis of scientific psychology, is an example of how difficult it is to relinquish this way of conceiving of ourselves. Alternatives to folk psychological theory (Churchland, 1979 and Stich, 1983) are hard to understand, and even harder to accept. How, for example, are we to redescribe action in terms other than those of folk psychology or some very similar ones? It is, perhaps, not insignificant that other disciplines concerned with human behaviour, like decision theory, psycho-analytic theory, and various philosophical theories of mind and action, are elaborations of the basic ideas of folk psychological theory. They all work with representational states that stand in causal relations to each other in a way that is connected to their semantic contents. However, whereas these features of folk psychology are important, they are not, I have argued, to be regarded as constitutive of the theoreticity of the body of information that is causally efficacious its production.

internal

5. Conclusion

In this chapter, I have examined two more substantial views of the theoreticity of folk psychology. They both model the theoreticity of the body of folk psychological information on the theoreticity of scientific theories. I have concentrated on the static features of scientific theories where possible, since we are concerned with Synchronic Theory Theory primarily. The aim has been to discover what is plausibly seen as making scientific theories *theories*, not what makes them *scientific* theories. It is accepted that, as well as there being a number of similarities between the body of folk psychological information and scientific theories, there are also a number of dissimilarities. This, however, neither detracts from the theoreticity of the body of folk psychological information, nor does it make the modelling uninformative.

According to what I call Traditional Theory Theory, a body of information is a theory if it is lawlike, posits abstract entities, and so on. I argued that a number of these constraints are far too strict to describe scientific theories at large. I therefore proposed a considerably more modest list of features. According to it, a body of information is a theory if it consists of a number of lawlike generalisations that contain terms that refer to entities and properties that explain some data, for example by being causally efficacious in the production of them, or by being related to them in some lawlike fashion, and the terms form a coherent, interrelated structure. We also saw that it provides a good model for the theoreticity of the body of folk psychological information.

The other contender, I dubbed Framework Theory Theory. According to it, the body of folk psychological information is a theory because it is similar to a paradigm, in the sense of being similar to a research tradition in science, rather than a specific theory. I go on to suggest that we reject the Framework Theory Theory because it seems to build too much into the theoreticity of folk psychology. It is an unnecessary strong claim to embrace in any foundational work. Consequently, we should embrace the modified traditional Theory Theory picture. It provides a good theory of the theoreticity of scientific theories, and it fits well with the body of folk psychological information as we know it. A consequence of this is, that Stich & Nichols' much more encompassing view of folk psychological theory will no longer count as a Theory Theory. It is a distinctive and interesting theory, but I think that the body of folk psychological information should not be regarded as a theory on that basis, for the reasons argued in chapter 2. Instead, we might call Stich & Nichols' theory of folk psychology 'Information Theory'. As such it is another theory of folk psychology, competing both with Theory Theory and Simulation Theory.

Chapter 4

Functionalism, Self-Attribution & Self-Knowledge

In the cognitive scientific as well as the philosophical community, the most popular account of people's understanding of mental-state language is the "theory of mind" theory, according to which naive speakers, even children, have a theory of mental states and understand mental words *solely* in terms of that theory. The most precise statement of this position is the philosophical doctrine of *functionalism*, which states that the crucial or defining feature of any type of mental state consists of its causal relations to (1) environmental or proximal inputs, (2) other types of mental states, and (3) behavioral outputs. (Goldman, 1993, p. 351)

The core of the functionalist strategy is the assumption that explanation of action or mental state through mention of beliefs, desires, emotions, etc. is causal. The approach is resolutely third personal. The Cartesian introspectionist error - the idea that from some direct confrontation with psychological items in our own case we learn their nature - is repudiated. (Heal, 1986, p. 45)

The main topics of this chapter are self-attribution and self-knowledge. As indicated in the introductory quotations, it is often assumed that Theory Theory is a functionalist theory and therefore committed to some functionalist account of self-attribution and self-knowledge (Heal, 1986; Goldman, 1993). Functionalism is sometimes assumed to hold that there is no difference in how we make attributions to self and others and, consequently, that we do not have any distinctive knowledge of our own psychological states. I therefore need to address the question of whether Theory Theory is a functionalist theory also. Consequently, I shall be concerned with three issues. Firstly, I shall examine the claim that Theory Theory is a functionalist theory, secondly, I shall consider what commitments Theory Theory has concerning how we attribute psychological states to ourselves, and thirdly, whether Theory Theory allows us to have some sort of distinctive knowledge of our own psychological states. This is the general structure of the chapter, but to do the issues justice, I must proceed in several steps.

There are many kinds of functionalism. Here I shall consider only metaphysical and semantic functionalism of the common sense variety, since this is the form of functionalism that Theory Theory is equated with. I shall argue that Theory Theory is not itself a functionalist theory, albeit compatible with either form discussed. However, even assuming that Heal is right that functionalist accounts of self-attribution are resolutely third personal, it is not clear that simply rejecting that Theory Theory is not a functionalist theory will do. For Theory Theory could still be committed to some kind of symmetric account of psychological attribution - that is, an account according to which attributions to self and others are based on the same evidence or grounds. An important consequence of such an account, is that it naturally leads to a rejection of the idea that we have distinctive kind of knowledge of our own minds. As Theory Theory has been presented so far, it would appear to be a symmetric account.

Embarking on the issue of self-attribution, I shall begin by posing a certain dilemma. Neither symmetric theories of psychological attribution, nor asymmetric ones appear to account satisfactorily for self-attribution. The former is deeply problematic, as we appear to have different grounds for first and third person attribution. The latter is difficult because an asymmetric account naturally gives rise to solipsism. If our grounds for psychological attribution are so different, how do we know that we are attributing the same states in the two cases? I take it that any reasonable account of self-attribution must satisfactorily navigate between these two extremes; including Theory Theory. So Theory Theory cannot simply be committed to a symmetric account of psychological attribution - indeed it is hard to see how functionalism can be. I will make it clear later, just how such an account constrains the account one can give of self-knowledge.

I then move on to consider some arguments to the effect that we don't have the kind of self-knowledge that we generally take ourselves to have of our conscious psychological states. These are taken from developmental and experimental psychology. They concern our direct access to the intentionality or representationality of our psychological states and our reasons. I conclude that these arguments indicate that we have less self-knowledge than we think we do.

I then turn to Theory Theory's commitments on these issues. I present a myth of the ontogenesis of self-attribution to outline these commitments. Theory Theory must place its account of self-attribution somewhere between symmetric and asymmetric accounts. What I present, is a version of how this occurs in developmental terms. The idea, though, can be extended to apply to a Synchronic Theory Theory. Once we have seen this, we can move on to self-knowledge to see whether such an account of self-attribution allows us to have a distinctive kind of knowledge of our own psychological states that we don't have of those of others. We will see that functionalism is not committed to attributional symmetry, and will consider two functionalist accounts of self-knowledge. I shall also consider Christopher Peacocke's account of self-knowledge. We will discover that prevalent accounts of self-knowledge, far from being dissimilar to that of the Theory Theory, are either fully or nearly compatible with it. This is not surprising, if any satisfactory account of self-knowledge cannot ascribe to either a full-blown symmetric or asymmetric account of psychological attribution. Allowing some symmetry between first person and third person attributions, does not commit one to denying that we have distinctive knowledge of our own minds. I conclude that Theory Theory advocates a view of self-attribution that allows it to be based on grounds different from those of attribution to others, and that we are entitled to knowledge of our own psychological states in a way that we are not with respect to those of others. I do not, however, go into details with this entitlement.

1. *Functionalism*

The two main varieties of functionalism are common sense functionalism - or *a priori* functionalism - and empirical functionalism - or psychofunctionalism (cf. Block, 1980). Theory Theory is normally only connected with the former. The best known proponent of this position is David Lewis (Lewis, 1966, 1972). According to him, psychological states are defined by the causal roles that they occupy (cf. the Ramsey-Carnap-Lewis theory of the meaning of theoretical terms). The causal role of a psychological state is its pattern of typical causes and effects. Causal roles are also known as functional roles. Typical causes of psychological states are states or events of the environment, other psychological states, and behaviour. Typical effects are behaviour and other psychological states. Folk psychological theory tells us just what causal roles the different psychological states occupy. Whereas for Lewis, a functional state is the occupant of a functional role, other functionalists, for example Putnam (1975b), prefer to view functional states as functional role-states. Nevertheless, the essence of the claim is the same: psychological states are identified in terms of functional roles. One way of describing functionalism is to say that according to it, psychological states are functional states picked out by some psychological theory. The version of functionalism that we will be concerned with says that this theory is folk psychological theory.

Theory Theory has been assumed to be committed to two forms of common sense functionalism: metaphysical functionalism and semantic functionalism. Metaphysical functionalism is the position that what it is to be a psychological state is to be a particular kind of functional state. Semantic functionalism is the idea that the meaning of mental state terms is defined according to the role such terms play in the theory in which they figure. I shall discuss each position separately, beginning with semantic functionalism.

2. Semantic Functionalism and Theory Theory

Semantic functionalism is a theory about the meaning of terms, more specifically, the meaning of psychological terms. These are treated as theoretical terms. Hence, common sense semantic functionalism can be regarded as a theory about the meaning of theoretical terms. We have already been introduced to this idea in chapter 3 - the Ramsey-Carnap-Lewis theory of the meaning of theoretical terms. We write the relevant theory *T* in a sentence: the postulate of *T*. We replace all theoretical terms *t*'s with unbound variables $x_1...x_n$, and, binding them with an existential quantifier, we get the Ramsey sentence of *T*:

$$\exists(x)\ T(x)$$

At this point, Lewis (1972) introduces the notion of a modified Ramsey sentence to get a unique realisation of *T*. Getting a unique realisation of *T* allows Lewis to identify functional states in terms of the occupants of causal roles, as opposed to a Putnamian functionalism where the functional roles would be identified in terms of the causal roles themselves. Introducing the notion of a Carnap sentence allows us to derive a meaning postulate from *T*. I will not go into detail with these technicalities. It is sufficient for our purposes to see that we end up with the following meaning postulate:

$$t = \text{the } x\ T(x)$$

In this sentence, theoretical terms are defined by the relation that they bear to each other and to other terms featuring in the theory. Let us see how this would work with folk psychological terms. For simplicity, let us regard the following statement as all the sentences of our folk psychological theory that involve 'belief':

> Beliefs are typically caused by perceptions and other beliefs, typically cause other beliefs, and, combined with desires, cause intention and action.

Reformulating, for ease and precision, 'beliefs' as 'instances of belief', and making the operations on this sentence that Lewis suggests, we end up with the following meaning postulate:

> Belief $=_{\mathrm{df}}$ the x_1 (instances of x_1 are caused by states of the environment and other instances of x_1, and instances of x_1 cause other instances of x_1, and combined with $x_{2,}$ instances of x_1 cause instances of x_3 and certain characteristic behaviours)

Notice, that in order to define our theoretical terms in terms of other terms, we have replaced states of perception with the environmental states that are assumed to cause the perceptions, and we talk of characteristic behaviours instead of actions. It need not be done quite this way. It is, perhaps, more plausible to keep perception in the sentence, and wait for states of the environment to be related to this definition by ways of the role that they play in defining perception.

We can now ask whether Theory Theory is committed to giving such an account of the meaning of folk psychological terms. As I have defined Theory Theory, it is certainly committed to such terms being theoretical terms. It is not, however, committed to any particular account of the meaning of theoretical terms. Instead of a Ramsey-Carnap-Lewis view, theory theorists can embrace the Kripke-Putnam view of the meaning of theoretical terms discussed in chapter 3. So, instead of understanding the meaning of terms in terms of definitions, one understands such meaning primarily in terms of causal links with the environment (Kripke, 1980; Putnam, 1973 & 1975a). As long as there are other satisfactory accounts of theoretical terms, theory theorists need not be semantic functionalists. Semantic functionalists need not be - although they frequently are - theory theorists. int or ext? Psychological terms could be defined in terms of a body of psychological knowledge that we all possess, but that is not what is causally efficacious in our folk psychological attributions. In other words, the definitions could be given in terms of an external account of folk psychology. Only when this is done in terms of an internal account, does the account count as a Theory Theory. Therefore, semantic functionalism and Theory Theory are separate positions.

3. Metaphysical Functionalism and Theory Theory

Metaphysical functionalism is metaphysical because it is a thesis about the nature of psychological states. Metaphysical common sense functionalism holds that the nature of psychological states is given by their causal roles as specified by folk psychological theory. This connects closely to semantic functionalism. Although it is not necessary to be a semantic functionalist to be a metaphysical functionalist - one can suppose that it is going to turn out to be an empirical truth that psychological states are how folk psychological theory says they are rather than it being analytic (see below) - metaphysical common sense functionalists happen to be semantic functionalists also.

According to Lewis (1972), the truth of metaphysical functionalism derives from the truth of semantic functionalism. This has the consequence that if there are psychological states, they must be as the theory says that they are. That is, the nature of psychological states is given *a priori* by the theory in which terms referring to them figure. There is no possible world in which we have psychological states, and folk psychological theory is not largely true of us. Either folk psychological theory is largely true, or no one has psychological states (Lewis, 1972, p. 213). Since what we mean by pain, say, is the state that plays the causal role specified by folk psychological theory, for anything to be pain it has to occupy this causal role, or very nearly occupy it. It has to be the state that is typically caused by bodily injury, that causes wincing, crying, and so on.

Although Theory Theory and metaphysical common sense functionalism are connected, they do not entail each other. Theory Theory is quite a strong empirical hypothesis about what is causally efficacious in our folk psychological attributions. As I stressed just above, it is an *internal* account of folk psychology. It is possible to be a metaphysical functionalist on the basis of an *external* account of folk psychology. Lewis (1994) seems to operate with such a view. He thinks that the principles of folk psychology that are explanatory of our psychological attributions do not simply boil down to the platitudes that he has argued are definitive of the meaning of psychological terms (1972). Knowledge of these principles is tacit (1994, p. 416). What is particularly interesting for us, is that it allows that one can be both a metaphysical common sense functionalist and a simulationist at the same time. It would be an unorthodox position, but not untenable as long as one's functionalism plays the role of an external account of folk psychology, and one's simulationism represents an internal account thereof. Metaphysical common sense functionalism *itself* is neutral with respect to whether folk psychological theory constitutes an internal or an external account of folk psychology. Since Theory Theory is an internal account of folk psychology, we can conclude that metaphysical common sense functionalism does not entail Theory Theory.

Metaphysical common sense functionalism is not implied by Theory Theory. Theory Theory is a theory about the mechanism(s) underlying our folk psychological practice. It is a theory of our beliefs about psychological states, not of the underlying nature of these states. More precisely, as a theory theorist, one is not committed to the view that either folk psychological theory is largely true, or we don't have psychological states. A theory theorist is free to hold both that folk psychological theory is false and that there are such things as psychological states. What a future psychology says our psychological states are like might be quite different from what folk psychology says they are, but this does not alter the fact that both theories refer to the same thing. Only, the former is true of these things, whereas the latter is false. On the other hand, if one is a semantic functionalist, one is committed to the view that if folk psychological theory is substantially false, we don't have any psychological states. One can be a theory theorist without being a common sense metaphysical functionalist and *vice versa*.

4. Functionalism, Self- Attribution & Self-Knowledge

Above we have seen that Theory Theory is not committed to either semantic or metaphysical functionalism. This means that one cannot use the commitment of Theory Theory to any of these theories, to argue that it is committed to a functionalist theory of self-knowledge. For example, Goldman (1993) identifies Theory Theory with semantic functionalism and then goes on to criticise this position on the basis that it does not provide a satisfactory account of how we attribute psychological states to ourselves.[39] He addresses three different forms of functionalism, only one of which seems to be what Heal calls a resolutely third personal account. However, as we have seen, this line of criticism doesn't touch Theory Theory. In fact, Goldman's' criticism relies on a flawed account of categorisation (Campbell & Bickhard, 1993), and functionalist theories of self-knowledge, themselves, do not appear to be seriously touched by it (Loar, 1993). We shall discuss some such accounts in section 8.

By highlighting the third person approach that she takes to be inherent in Theory Theory, Heal (1986) points towards a further consequence of holding a symmetric account of psychological attribution. If our self-attributions are based on the same grounds as our attributions to others, how can the knowledge that we have of our own psychological states be any different from that which we have of others' psychological states? So the issue of self-knowledge is intimately connected to that of self-attribution. The issue now becomes: even if we have rejected the idea that Theory Theory is a functionalist theory, there may still be elements in the Theory Theory that are such that it can only licence a symmetric account of psychological attribution. This idea would appear to be functionalist in spirit, although amounting to neither metaphysical nor semantic common sense functionalism.

Our psychological attributions to others seem to be based on an inference from the observable causes and effects of psychological states. We perceive that people say or do certain things, that they are placed in such-and-such an environment under such-and-such circumstances, and having ascertained this we can apply our folk psychological theory to generate psychological attributions. One might

[39] On p. 370, Goldman says that "commitment to a TT [Theory Theory, ed.] approach does not necessarily imply commitment to RF [Representational Functionalism, ed.] in the mental domain; nor would evidential corroboration of a TT approach necessarily corroborate RF." This seems to indicate that he is aware of the difference between Theory Theory and functionalism. Nevertheless, his whole paper revolves around rejecting RF such as to throw the viability of Theory Theory into doubt. RF is a variation of semantic functionalism. In order to stress the way in which a "cognizer ... represents mental words" (p. 352), rather than knowledge of the meaning of psychological terms, Goldman uses the term 'RF'.

recast this in functionalist terms. The causal roles of psychological states play a crucial role in us being able to attribute such states. There are two parts to this. First, there is the idea that our psychological attributions to others are based on some kind of inference that takes as its starting point observable causes and effects of these states, such as behaviour and environmental factors. Among all but Wittgensteinians, this is fairly uncontroversial. Second comes the idea that we need to deploy a theory in our inferential reasoning from what we can observe to the psychological states themselves. This *is* controversial, and quite specific to the Theory Theory.

The worry, then, is that Theory Theory is committed to an account, in which all our psychological attributions are based on the observable causes and effects of psychological states. Such an account seems third personal in nature, as Heal pointed out. However, as I presented Theory Theory in chapter 1, it is a theory about folk psychology defined as our practice of attributing psychological states to *everybody*, including ourselves. This means that we cannot understand Theory Theory simply as an account of third person attribution, self-attribution being an entirely separate matter. Therefore, Theory Theory appears committed to a symmetric account of folk psychological attribution. But if we accept this, then we also seem forced to accept that we cannot claim to have any distinctive knowledge of our own psychological states as opposed to those of others. For how could we possibly justify such a claim if we base our attributions on data of the same sort?

I cannot answer this question yet. We must first address a number of other issues. First of all, I want to present a dilemma for theories about self-attribution and self-knowledge. This shows just how such theories must be framed. Then, in section 6, I will turn to arguments to the effect that we have much less self-knowledge than is normally assumed. The extent of our self-knowledge is a crucial datum we need in order to theorise about it. A satisfactory account of self-knowledge must explain just why our self-knowledge is so limited. Once I have discussed these issues, I can return to discuss Theory Theory's commitments concerning self-attribution and self-knowledge (section 7 & 8).

5. The Dilemma

We have seen that embracing a symmetric account of psychological attribution has the consequence that we do not have any distinctive knowledge of our own minds. However, embracing an asymmetric view also has certain important consequences. It leads to solipsism. As such one might wonder whether we can have self-*knowledge* if our concepts themselves are first-personal in nature. I shall leave this question open and concentrate on solipsism. But let us first look at the problems facing a symmetric account.

One of the best known symmetric accounts of psychological attribution is behaviourism (Ryle, 1949). To be in a psychological state, is to be disposed to act in a particular way. Psychological states are dispositions to act. However, the way in which psychological states are attributed is on the basis of behaviour. Psychological attribution is symmetric, although being placed as close to ourselves as we are, we gain a certain expertise in self-attributing states, that we don't have with respect of others.[40] This might give rise to the illusion that we know ourselves in a way that we don't know others. The problem with this view and all symmetric views of self-attribution is that it just seems self-evident that in some cases, at the very least, we do self-attribute psychological states on a basis on which we cannot attribute psychological states to others. When, for example, I am sitting quietly at my desk with my eyes closed musing about my summer holiday, I have no behaviour to go on. Even when I am not moving or perceiving, I can attribute thoughts to myself.

It seems that we are placed quite differently with respect to others than with respect to ourselves. It is not simply that I am the occupant of my body. It is also the case that I am capable of being aware of my sensations, feelings, and thoughts in a way very different from that in which I can become aware of the sensations, feelings, and thoughts of others. I feel my own movements while I make them. I am immediately aware of what I see, hear, smell, and taste. With respect of

[40] Ryle says that one can listen in on one's own silent soliloquies (p. 162), but the difference in the kind of knowledge that we have of ourselves and that we have of others, is still a difference "of degree, not of kind. The superiority of the speaker's knowledge of what he is doing over that of the listener does not indicate that he has Privileged Access to facts of a type inevitably inaccessible to the listener, but only that he is in a very good position to know what the listener is often in a very poor position to know." (p. 171).

others, I have to check that they are located in the right position, that their eyes are open, their noses unblocked, their hand stretched out, etc. More poignantly, when others are in pain, I do not feel their pain although I may sympathise, or even empathise, with them. When I am in pain, I just *know* that I am in pain. In most cases, I also know just what beliefs, desires, hopes, fears, and so on, I have. I need not observe my environment or my behaviour.[41] I may need to reflect on the matter if, for example, I am asked whether I believe that it is possible that God should be both omnipotent and omniscient. However, if the question concerns my present conscious thoughts, I do not generally need to do so.

In general, it is supposed that we have access to our own psychological states that is immediate, non-inferential, privileged, authoritative, and immune to error through mis-identification. All of these characteristics stand in contrast to those of our knowledge of third personal psychological states. We can sometimes attribute psychological states immediately, without paying attention to anything, but what goes on in our minds. We need not always infer consciously? what we think and feel. This is a privilege that we have, for no one else can attribute psychological states to me on those grounds. Such self-attributions have the status of self-knowledge. We are assumed to be authoritative with respect our self-attributions. If I say that I thought *p* then, in the absence of strong countervailing evidence (me acting as if I thought ~*p*, for example), nobody will presume to argue with me on that point. When others are in doubt as to what I think, they ask me to

[41] This is not to say that I'm impervious to the environment when I attribute myself psychological states. From my point of view, my beliefs are accurate representations of my environment. Therefore, it makes sense for me to examine my environment to figure out what I believe (cf. Evans, 1982). Thus, if I ask myself whether I believe it is going to rain, one way in which I might go about answering this question is by looking at the sky, feeling the humidity in the air, look up the weather forecast, and so on. Although there is a sense in which I don't really have a fixed belief at the time of the question, informing myself that I believe that it will rain, is neither beside the point, nor false. Note, however, that when I observe the environment in order to figure out my beliefs, I am still very differently positioned with respect of myself than with respect of others. In my own case, I do not need to consider both the environment and the position of my sense organs with respect of it (are my eyes opened, pointed in the right direction, and so on). I am directly aware of what I perceive. Not so, in the case of other people. Here I need to make sure that they are in the right perceptual relation to their environment, when I attribute beliefs to them. To make sure that you saw what I saw, I need to make sure that you were looking in the right direction with your eyes open, and so on. So, my use of environmental information must be supplemented by information about the perceptual location of the subject in the case of others, whereas in my own case, there is no such need. I am, as it were, immersed in my own point of view. This constitutes an important difference between how I other- and self-attribute.

tell them. If they are in disagreement about what I think, and come to me to tell them, they may require me to argue for why I think something, but not usually for why I think that I think what I think. Lastly, there are certain errors that I cannot perpetrate with respect of my privileged self-attributions. I cannot mistake who it is that has the states that I attribute to myself on a non-observational basis (on the usual reading of observation as observation of external affairs). I cannot think "someone believes that it is raining, but who is it?" when the belief attribution is non-observationally based (Shoemaker, 1968).

It appears that only the psychological states that we are conscious of having, are states that we can self-attribute non-observationally. Since there may be psychological states that are not conscious, it is often useful to attribute psychological states to oneself on the basis of what one says or does. It is likely to furnish one with much knowledge of oneself. The idea of there being unconscious psychological states is now widely accepted, even in its psycho-analytic formulation (Freud, 1915/1957). Unconscious psychological states include repressed emotions and ideas, but also unrepressed unconscious ideas. So-called tacit knowledge states can also be included as unconscious psychological states. Tacit knowledge is attributed to people to explain a capacity that they have, but the underpinnings of which they have no explicit knowledge of, for example, knowledge of grammar and visual parameters (cf. chapter 5). There are also implicit knowledge states. Such knowledge is involved in tasks "that are [...] overlearned, routine, or of minimal interest" (Smith & Miller, 1978, p. 361). In these cases, subjects are not conscious of the knowledge that is causally efficacious in the particular task. I cannot here go into the similarities and differences between tacit and implicit knowledge, but will discuss some such issues in chapter 5. One might also think that there are less extraordinary psychological states that we are also unaware of. Take Austen's *Emma.* Emma infers that she is in love with Mr. Knightly because of her very violent reaction to the suggestion that he might be about to marry somebody else. She comes to realise her love not, as is more usual, by directly feeling it, but by analysing her emotional reaction to a particular event. On the face of it, this appears to be a case of Emma becoming conscious of her feelings. It cannot be ruled out, however, that the issue concerns awareness and not consciousness as such.

To make what has become rather a long story shorter, there seems to be little future in the idea that we base our psychological attributions on the very same evidence in all cases. There appears to be an asymmetry between certain cases of first person attribution and all other attribution. This is the one horn of the dilemma.

The other horn of the dilemma is that it seems equally problematic to hold an asymmetric account. The problem is that if we attribute psychological states to ourselves on a basis different from that on which we attribute psychological states to others, how do we know that we are attributing the same thing in the two cases? More precisely, how do we know that the psychological states that we self-attribute non-inferentially are the same kind of states as the psychological states that we attribute inferentially? Given that I self-attribute psychological states on the basis of *feeling* certain things or *being aware* of certain things, how can attributions of psychological states have the same meaning in the cases where I attribute them in the absence of being (directly) aware of or feeling these states? In the words of Thomas Nagel, it leaves quite open the possibility that "mental attributions do not have the same sense in the first person as in the third" (Nagel, 1986, p. 20). This problem is sometimes known as "solipsism" (Strawson, 1959, p. 87). Thus, if one accepts asymmetry between first person and third person attributions, one must come up with an account of what makes these attributions, attributions of the same kind. We need to know exactly why it is that what we apply in the two different situations are unified concepts of psychological states. A popular solution is to build it into the possession conditions of psychological concepts that they are applicable under both circumstances, and only when one knows that they are applicable under these two different circumstances, does one possess the concepts (Strawson, 1959; Peacocke, 1992).

The solution, then, must lie between the horns, as it were. Self-attribution is asymmetric to third person attribution in certain respects, but not in others. I take it that this is the shape that any satisfactory account of self-attribution and self-knowledge must take. Later, we shall see that Theory Theory can navigate successfully between a completely symmetric and a completely asymmetric account. Before doing this, we must first examine (some of) the limits to self-knowledge.

6. The Limits of Self-Knowledge

I shall look at two criticisms of the idea that we have direct access to our psychological states; Alison Gopnik's and Richard Nisbett, Lee Ross & Timothy Wilson's (Gopnik, 1993a; Nisbett & Ross, 1980; Nisbett & Wilson, 1977). These criticisms suggest that we do not have privileged access to either the intentional or the causal aspect of our psychological states.

According to Gopnik (1993a), we have much less direct access to our psychological states than we think we do. In particular, the intentional aspect of these states is not directly given in them, but is an inferred characteristic. She argues this specifically with respect of belief, but there are very good reasons to suppose that she believes the same to hold for all other intentional psychological states (Gopnik & Slaughter, 1991; Gopnik & Meltzoff, 1997). As we saw in chapter 1, young children fail to attribute psychological states to themselves and others that are at a variance with reality as they see it now. This is true even in situations where they, themselves, have claimed to hold the opposite belief - that there were smarties and not pencils in the container. According to Gopnik, experiments such as these support the Theory Theory. Young children fail to pass the false belief task because they are in the grips of a proto-folk psychological theory - what Wellman (1990) calls 'the Copy Theory of Belief' - according to which beliefs reflect reality and, consequently, cannot be false. Hence, even when they have made a different report earlier - "I think that there are smarties in the box" - they will claim the opposite, once the true contents of the box have been revealed - "I thought that there were pencils in the box". The idea is that young children do not yet have a concept of the *intentionality* of beliefs. They believe that all their beliefs correspond to reality and hence, when there is a conflict, report their former beliefs as conforming to reality as they see it at the present.[42]

[42] It is quite possible that this is specific to cases where what is at issue is a belief based on an expectation versus a later belief based on perception, as opposed to an earlier and later belief both based on perceptions. In the former case, there is no

There is a puzzle here. It doesn't seem that children have a memory problem, and hence simply cannot remember what they thought earlier. Gopnik has conducted many tests of comparable circumstances, for example moving things around and questioning children about where ~~they believed~~ the things were before (1993a). In these tasks children elicit few problems recalling their prior beliefs. However, if we assume that they do remember their earlier beliefs, then we end up with the implausible situation in which the children remember their old belief and then think "no, that can't be right because it doesn't fit with how things are now". But this, Gopnik claims, is the wrong way to look at things. For we are assuming that children either have direct access to their beliefs or that they have a problem remembering them. However, the situation is more sinister - forgive the expression - than that. Beliefs aren't simply given to children. They are theoretical entities constructed by them for predictive and explanatory purposes. This is not to say that they do not have psychological states, but simply that they do not experience them as the kind of entities that they are later to understand them as; as we understand them. Gopnik draws a direct parallel between the way in which our knowledge of folk psychology develops, and the way in which we - as adults - have access to our psychological states. Gopnik says "we may well be equipped to detect certain kinds of internal cognitive activity in a vague and unspecified way, what we might call "the Cartesian buzz"" (p. 11). Nevertheless, psychological states are not directly given to us *as such*, we come to believe that they are so because we gain expertise from self-attributing such states. Expertise, Gopnik says, often gives rise to the illusion that one directly apprehends something that one, ~~in~~ fact, only infers the existence of.

There is one very important objection to Gopnik's conclusion that the intentionality of our psychological states is not directly given to us. It may simply be the case that children don't have a concept of belief before the age of 4, give or take some months. Shoemaker (1993), for example, claims that one cannot self-ascribe proper belief states before one has the concept of belief, and for all that Gopnik has shown, we cannot exclude that the lack of the concept of belief is what

perception to hold on to, to justify one's prior belief. This may be one reason that it is reinterpreted. This is still consistent with Gopnik, because it seems to be the idea that beliefs are directly linked to reality that prompts the reinterpretation. In other words, it is a theoretical assumption that prompts the reinterpretation.

is at issue. To put the matter differently, the fact that one needs to possess the relevant concepts in order to be able to self-attribute psychological states, has no consequences for what, if anything, is directly given to us in our conscious psychological experience. The situation here could be absolutely standard, and in no way different from ordinary concept acquisition. I cannot self-attribute the belief that endorphins are released during exercise, before I have acquired the concept 'endorphin'. But this is just the relatively boring point that one cannot ascribe certain properties to things if one doesn't possess concepts of those properties. Not having a concept of something, has no consequences for the direct availability or observability of that something: (van Fraassen, 1980, p. 15)

> It is also important here not to confuse *observing* (an entity, such as a thing, event, or process) and *observing that* (something or other is the case). Suppose one of the Stone Age people recently found in the Philippines is shown a tennis ball or a car crash. From his behaviour, we see that he has noticed them, for example, he picks up the ball and throws it. But he has not seen *that* it is a tennis ball, or *that* some event is a car crash, for he does not even have those concepts. He cannot get that information through perception: he would first have to learn a great deal. To say that he does not see the same things and events as we do, however is just silly; it is a pun which trades on the ambiguity between seeing and seeing that.

The fact that someone doesn't just see a tennis ball *as* a tennis ball, does not mean that she doesn't see the tennis ball, nor does it mean that, once the concept 'tennis ball' is acquired, seeing a tennis ball does not play an important justificatory role in, for example, her belief that there is a tennis ball on the lawn. To return to the issue at hand, the fact that in order to understand our beliefs as intentional states, we need to possess the concept 'belief', does not, by itself, imply that the intentionality of our psychological states is not directly given to us. Nor does it mean that this givenness cannot play a justificatory role in our self-ascriptions, such that we can regard them as constituting knowledge. In short, children need not be understood as *inferring* that their beliefs are intentional states. ?

We have to be careful here, however, lest we should bar the possibility of misapplying concepts. That is, we cannot simply say of all misapplications of concepts that they are due to the subject applying them not possessing the concepts. This would make our claim vacuous and unfalsifiable. Say that we allow for occasional misapplication - can we also allow that there is systematic misapplication? The case of children failing the false belief task must be understood as systematic misapplication, if we allow that they possess the concept of 'belief'. As Gopnik (1993b) points out, children use the terms 'think' and 'know' (the child's equivalent of philosophers' 'belief') appropriately, and appear to understand them. They also "know that thoughts are different from things and that one person can have a thought about something while another person may not have a thought about the thing." Do we still want to deny that they possess the concept of 'belief', but are systematically misapplying it in just one kind of circumstance - that of false belief? I think this is a fair and important question to raise in this context. Unfortunately it opens a can of worms that is impossible to deal with here. Instead, I will focus a defence of the idea that we do not have direct access to the representationality of our psychological states on a comparison between the development of children's understanding of belief, on the one hand, and desire and perception, on the other.

As indicated in chapter 1, children develop a representational understanding of perception and desire much in the way that the develop an understanding of belief (e.g. Gopnik & Meltzoff, 1997). At an early stage of development, children seem to appreciate the content of their representations, but not the fact that they are representations. Before the age of 2.5-3, children don't fully grasp that *esse* isn't *percipi*. That is, they have problems understanding visual perspective and occlusion, for example. More poignantly, children fail "changed desire tasks". In an experiment, Gopnik & Slaughter (1991) presented children with two apparently equally desirable items, books, among others. Children were asked which one they wanted, after which the experimenter read them the chosen book. Afterwards, they were asked which book they wanted to have read to them at the beginning of the experiment. Three year olds invariably claim that they had wanted the other book to be read to them. The situation is similar when children are queried about their desire before and after eating snacks. Before they will profess that they desire the snacks, after they will deny it. It seems, then, that there is a reinterpretation of past desires in the light of present ones. It seems that children fail to understand that their desires may change over time. This, in turn, can be put down to a general inability to appreciate the representational aspect of desire. The reason is that children appear not to have problems reporting their present desires. A problem only arises once you have a present desire that is inconsistent with a past desire - then the past desire gets to be misreported. This lends support to the view that the representationality, as opposed to the content, of psychological states is not directly given to children. Because they are not immediately presented with the representationality of psychological states, they become confused in certain situations and attribute themselves states of the right type (desire, say) but with the wrong *content*.

the content, or that it is a ? representa

I think that experiments such as the above lend support to the idea that the representationality of our psychological states is not directly given to us. We infer it. Children appear to have problems understanding the intentional aspect of their psychological states. This is explicable in terms of this aspect not being directly given to them. It seems plausible to suppose that we are just in the same situation. It is not the case that with development, we come to be directly presented with the representationality of psychological states. But, knowing that they are representational, we tend to behave accordingly. It is possible here to reiterate the concept argument. I think, however, that it is unlikely to be a coincidence that it is the representationality of psychological states that creates these big problems for children. To assume that it is the conceptualisation of representational states that is at issue does not explain the data as well as the idea that this representationality is not directly given to us, whereas other aspects of our psychological states are. This is not, I know, a knock-down argument in favour of the Gopnikian idea that the intentionality of our psychological states is not directly presented to us in our experience of our psychological states. Nevertheless, I think that there are good reasons, and I hope to have provided some, in favour of this view. In what follows, I shall assume that the evidence supports the Gopnikian idea. Let me stress, however, that what I take from Gopnik is simply the view that the representationality of our psychological states is not directly given to us, not something stronger like aspects of our psychological states not being directly given to us at all.

Nisbett, Ross, and Wilson have provided strong evidence that we have quite restricted direct access to our reasons for doing things (Nisbett & Ross, 1980; Nisbett & Wilson, 1977; Wilson, 1985). The evidence in question, is that of experimental subject's verbal reports on their reasons for judging or acting in particular ways, within the experimental set-up. The accuracy of the reports is calculated by so-called 'objective measures', primarily a mixture of non-verbal behaviour and verbal reports concerning current psychological states.[43]

In their seminal paper, "Telling More Than We Can Know: Verbal Reports on Mental Processes", Nisbett & Wilson argue that a number of different experiments show that subjects have "little or no direct introspective access to higher order cognitive processes" (p. 231). I take it that what they mean is not that we don't have direct access to psychological *causes* as such, but that we don't have such access to the causal *relations* of psychological states.[44] Psychological states do not carry their causal history on their sleeves. Let us have a look at one of their experiments.

The Position Effect. In this experiment, subjects are asked to choose the best quality token of a number of tokens of the same type displayed in a row, for example stockings. All the tokens are, in fact, indistinguishable with respect of quality. Thus, the stockings exhibited will be of the same make, style, with no apparent faults, and so on. It turns out that the right-most token is preferred to the others by a factor of almost four to one. When asked to justify their choices, subjects never refer to the position of the chosen item, but only to its quality. As a matter of fact, when queried, subjects deny that the particular position of the item *vis-à-vis* the other items influenced their choice.

[43] For criticisms of this assumption, see White (1988). For defence of behavioural measures, see Wilson (1985).

[44] With the notable exceptions of C. Ducasse (1926) and David Armstrong (1993), few philosophers believe that there is any direct access to the causal relation. Most are good Humeans in believing that causes are *inferred* from observed constant conjunction. For Nisbett & Wilson's article to present a threat, they cannot have in mind that we do not have direct access to psychological *causes*.

The conclusion that Nisbett & Wilson draw from this, is that the subjects had no direct access to their reasons for choosing the item that they did. Since the items were indistinguishable as to quality, it is unlikely that subjects really did perceive differences in quality. What seemed to determine quite a number of subjects' choices was the position, since the probability of random choice through the row would lead to the choosing of the rightmost item much less frequently than was, in fact, the case. Hence, what seemed to determine subjects' reports was not what state was causally efficacious in the production of the relevant action, but what best made sense of that action. Given the instructions, the stocking being of superior quality would make the most sense of the choosing of that stocking. According to Nisbett & Wilson, subjects have a theory of what counts as good reasons for thinking and acting in particular ways and they believe that people have good reasons for acting and thinking as they do, *ceteris paribus*. It is knowledge of this theory that is causally efficacious in the production of their reports. The rightmost bias is due to a "shopping around" habit. Shopping around implies withholding choice until all items are examined. This would explain the right hand bias, since most people evaluate items in a linear display left to right. Presumably, this habit is unconscious and hence the subjects self-attribute psychological states that they are not in.

Ian Ravenscroft has a more elaborate story to tell about the position effect. According to him, the causally efficacious states were consciously inaccessible and therefore, as most people would agree, not introspectible (Ravenscroft, MS). The position effect is caused by a particular *implementation* of the relatively high level psychological function of decision making. Psychological causation of behaviour consists of a hierarchy of instructions filtering from the high level decision to the low level motor implementation of the relevant behaviour. The implementing of a decision to act is a low level psychological activity - low level because it does not involve the abstract representations we standardly find in decisions and because it is causally more proximal to behaviour. Low level processes and states are *not* consciously accessible, only high level processes or states are. In terms of representations, outcomes of decision making processes are something like "go to the kitchen", whereas the implementation of this decision at a lower level is something like "activate muscle M to degree D" (p. 14). The same principle applies in the case of the position effect. Deciding to choose at random might not always lead to picking at random, since low level processes might favour a particular movement, the extension and grasping of the right hand, hence of rightmost items in a display (p. 17). So, according to Ravenscroft, the position effect is explicable by subjects indeed noticing that the items were of similar quality and deciding to choose at random. It just so happens that picking out objects at random from a linear display going from left to right, involves a motor reflex that is biased towards right-most items. Nevertheless, this fails to explain why subjects do not report their decision to choose at random. Deciding to choose at random is a high level psychological process and, as such, should be consciously accessible.

Let us return to Nisbett, Ross, and Wilson. I believe that their view is open to a number of interpretations. I think the following is the best. Subjects do not have conscious access to the causal properties of their psychological states under that description. Although they might have direct access to some of their psychological states, they must infer how these are causally related to one another. They infer this on the basis of knowledge of some theory or body of information about what it makes best sense to think and do under the circumstances. They may even do so, when they have memories that should be revelatory of the real reason for their actions: (Nisbett & Ross, 1980, p. 248)

> We propose that when people are asked to report how a particular stimulus influenced a particular response, they do so not by consulting a memory of the mediating process, but by applying or generating causal theories about the effects of that type of stimulus on that type of response. They simply make judgements, in other words, about how plausible it is that the stimulus would have influenced the response. These plausibility judgments exist prior to, or at least independently of, any actual contact with the particular stimulus embedded in a particular complex stimulus configuration.

Further experiments have shown that observers of experimental subjects make the same judgements as the subjects themselves, when asked to explain why subjects chose as they did. According to Nisbett

& Ross, these reports were "so strongly correlated for each of the judgements that it seems highly unlikely that subjects and observers could possibly have arrived at these reports by different means." (p. 250). Nisbett, Ross, and Wilson, then, seem to advocate a largely symmetric account with respect of attribution of explanatory psychological states, or reasons. Although we no doubt have direct access to *some* of the aspects of *some* of our psychological states, their causal relation are not amongst them.

So, what Nisbett, Ross, and Wilson's work makes clear to us, is that we don't have direct access to our reasons for doing things *as* the reasons for which we did those things, despite the fact that we may have access to these reasons as beliefs, desires, and so on. It is consistent with this view that in some cases the reasons that cause an agent to act in a particular way, are so salient to her that she will naturally regard those as being her reasons. However, even when there is a high probability that some belief-desire pair constituted her reason to act, by being the most salient one and the one most proximal (time wise) to the action, she might ignore such a pair in favour of some account that makes better sense of her action.

A consequence I want to draw from Nisbett, Ross, and Wilson's work, concerns the causal roles of our psychological states. The idea is that if we don't have any direct access to the causal relations of our psychological states, we don't have direct access to their causal roles either. This is due to the fact that we would normally infer causal roles on the basis of causal relations. However, since we don't have any direct access to these relations, our knowledge of them is based on an inference. Consequently, our knowledge of the causal roles of psychological states is doubly indirect, since it is based on an inference from causal relations that are themselves inferred. Presumably nobody ever wanted to claim that we have direct access to causal *roles*. It seems obvious that we infer causal roles from causal relations by considering which are most typical. However, it seems natural to claim that we have direct access to *these* relations. If this were true, there would be a sense in which we had relatively direct access to the causal roles of our psychological states, because these roles would have been inferred from experienced causal relations. However, Nisbett, Ross, and Wilson make us doubt that this is true. They make, I think, a convincing case for the idea that we don't have direct access to the causal relations that our psychological states stand in. This has the consequence that our access to causal roles is completely inferential. The means that we do not have any distinctive knowledge of why we do the things that we do, as opposed to why other people do what they do. However, since we do have knowledge of a number of our own psychological states, we are in a better position than others to determine which of our beliefs and desires were causally efficacious at any one time. Nisbett, Ross, and Wilson's experiments place a definite limit on our entitlement to self-knowledge. They do not, however, eliminate it.

Let me summarise what we have found in this section. We do not have direct access to the representationality and the causal roles of our conscious psychological states. When we self-ascribe psychological states, then, this is not done solely on the basis of awareness of these states. However, this does not impugn the fact that we do have some self-knowledge. It does, however, exclude that we have any kind of distinctive knowledge of what our reasons for doing things are, compared to the knowledge that we can have of others' reasons for doing things.

7. The Myth of the Ontogenesis of Self-Attribution

We can now turn to Theory Theory's commitments concerning self-attribution and self-knowledge. My aim is to show that a Theory Theory account is not incompatible with certain theories of self-knowledge. In short, I want to show that Heal's charge is unjustified. Before doing that, I will outline Theory Theory's commitments on these issues in terms of a myth. The Myth of the Ontogenesis of Self-Attribution. This is meant to recall the Myth of Our Rylean Ancestors - the first version of Theory Theory (Sellars, 1963, p. 178). I call it a myth because I want to indicate that it may be mistaken, but I don't mean to imply that it *is* false, nor that important parts of it are false. In fact, I rely on the myth being largely true. The purpose of the myth is to outline the development of psychological attribution in a way that clarifies why such attribution is neither fully symmetric nor fully asymmetric. Being consonant with at least some of the data from child psychology, it should have some empirical plausibility. The myth is not meant to illustrate a view a theory theorist must inevitably take on the development of psychological attribution. It is an empirically viable suggestion very much compatible with the Theory Theory. Keeping this in mind as a prototypical Diachronic Theory Theory, we can move on to determine the relative compatibility of Synchronic Theory Theory and theories of self-knowledge.

This is the myth. In many ways, infants are very different from us. Not only are they very small, pink, and loud, but they don't see very well, they cannot speak, their episodic memory is undeveloped, they appear not to reason much, if at all, and their interaction with their environment is severely limited. A host of cognitive abilities that normal adults possess, take years for the infant to develop. Infants don't appear to have concepts of psychological states, but despite arguments from philosophers such as Davidson (1975) to the effect that to have thoughts one needs to have the concepts of such thoughts, there are pretty convincing considerations for the mindfulness of young children.

First, there is the problem of how to explain what goes on in children's minds, if anything at all, before they get to have psychological concepts. If we assume that children only have a concept of belief once they are capable of passing the standard false belief test, they will be around four by that time. What goes on in their minds that is such that gaining a concept of belief will give them beliefs? I think you will agree that this is quite a hard question for a Davidsonian to answer. This is compounded by the fact that children clearly interact with their environment as if they have beliefs before the age of four. They also have rudimentary language at that stage and will talk of what they think, want, and feel. They are not easily regarded as automata reacting to their environment in pre-programmed ways. And were we, despite our better judgement, persuaded to regard them in this fashion, there would still be the problem of how genuine beliefs arise out of an instinctive reaction pattern, and how getting the relevant concept is definitive of such development. Nothing we know now of ontogenetic development, gives us any idea of how this could be.

Taking the view that children do have psychological states before they gain psychological concepts, is more natural and puts us in a better explanatory position. Children may, at first, experience a limited range of psychological states. For example, it may take some time to develop the capacity to feel such things as aesthetic pleasure, glee, and existential angst. Nevertheless, there is little reason to deny that they have beliefs and desires, and experience such emotions as pleasure, attachment, longing, anger, and upset, among others. Although young children only seem capable of certain thoughts and feelings, they are nevertheless disposed to feel the full range. They have conscious psychological states.

If we assume that we have some direct access to aspects of *our* psychological states, there are good reasons to suppose that children have some direct access to *their* conscious psychological states. Children talk about their psychological states, for example. It seems reasonable to suppose that although their capacity to access their own psychological states develops, children do enjoy some direct access to their own psychological states. Through interaction with others and considerable further experience, children come to conceptualise their psychological states as psychological states. This process takes years, and in those years the rudiments of folk psychological theory are acquired. Initially, children are aware of the phenomenology that accompanies many, or perhaps all, of their psychological states and experiences. When they first feel anger or joy, they do not know that this is what they feel, but they are aware of feeling something. The former feeling brings displeasure and the latter pleasure, and they strive to avoid the former and achieve the latter. Children are also aware of the various ways in which their psychological states represent the world. We must not project back into early childhood an adult, and philosophically sophisticated, understanding of such states. Children are aware of the world as being in one way or another, but this is not to say that the representationality or intentionality of their psychological states is directly given to them in their awareness of such states. There are at least two ways in which folk psychological theory contributes to our understanding of psychological states. It tells us that psychological states are representational states as well as what kind of representational states they are. That is, it tells us what direction of fit, for example, any given psychological state has. It does so through telling us of its causal role - what typically causes it and what it typically causes.

empiricism?

This way of putting things is, of course, somewhat misleading. It makes it sound as if we receive a folk psychological theory at some point and then set about applying it. Some theory theorists may believe that some or all of the core of folk psychological knowledge is innate, but it is certainly not necessary for them to do so. They can regard early development as an acquisition of knowledge about psychological states, among other things. Children *learn* that their psychological states are representations with a particular direction of fit. So whereas they might initially have attributed goodness to objects in the world - mother's breast is good, chocolate is delicious - they come to understand that this goodness is a projection of the satisfaction of their own desire. The goodness of chocolate comes to be ? understood in terms of a desire for chocolate that they have, not in terms of some intrinsic property of chocolate. With beliefs the situation is different. Children may at first simply regard their beliefs as yielding the world as it is - the world is directly given to them. Later, they come to understand that beliefs *aim* to represent the world as it is. However, it is in the very nature of representations that they can misrepresent. Hence, although beliefs aim at truth, they sometimes fail to capture it. Psychological states present themselves to us by presenting us with a content and an attitude towards that content. However, we need to know about the causal roles of such attitudes - beliefs tend to be caused by perceptions, other beliefs, etc. - in order to understand what is given to us, as a particular psychological state. Otherwise, we cannot grasp the significance of it. Furthermore, the content that we are presented with, is something that we learn represents the world in a particular light, all depending on the psychological mode. Learning these two aspects amounts to learning the rudiments of folk psychological theory.

Understanding the causal role of a psychological state is part and parcel of understanding what that psychological state is. Understanding that something is a belief implies understanding that it is the kind of state that typically is caused by other beliefs and the environment, and that tends to give rise to other beliefs, and so on. This is, of course, exactly the kind of information that we acquire when we acquire folk psychological theory. There is a strong and a weaker interpretation of this. On the strong one, what allows us to individuate some psychological state as a particular psychological state, is that we know its causal role. On the weaker interpretation, knowing the causal role of a psychological state plays a crucial role in understanding what kind of state it is, even if one is in a position to individuate it in the absence of that information. The stronger interpretation lends itself to semantic functionalism, the weaker doesn't.

This, then, is the myth. We can understand the initial state of self-awareness that young children have, as what is directly given to us in our awareness of our conscious psychological states. With knowledge of folk psychological theory and practice in applying it, we come to believe that we had direct access to psychological states as particular representations of the world with particular causal powers. Here's why the myth captures the commitments of the Theory Theory. On the one hand, it allows for some direct access to our own psychological states, but on the other it shows us that such access is limited. It is through the acquisition of knowledge of folk psychological theory, that we learn to see such states as fully fledged psychological states - representations with particular directions of fit, and with particular causal powers. We cannot acquire such a theory simply by sitting around reflecting on our thoughts. We normally do so through interaction with other people, by coming to understand that we are all subjects of psychological states. It is, perhaps, possible to acquire such a theory alone, but that could only be done by observing oneself as one would observe others - see what they (I) do and say. It is only through an understanding that psychological states manifest themselves both directly in thought and indirectly in behaviour, that we gain a full understanding of the nature of these states. What is clearly necessary is the ability to look at oneself from two different perspectives - from the inside and from the outside. This is not to be behaviourist and claim that all thought must be linked to dispositions to behaviour. All that is required is that in order to understand what beliefs, desires, hopes, and so on, are, we need to see how some such states are linked to behaviour. Other instances of that psychological kind need not necessarily be linked up with behaviour for us to comprehend them. Discrepancies between the environment and thoughts that one most typically connects with observation of others, can be achieved by a comparison between different psychological states at different times. Nevertheless, although it may be possible for a person to acquire folk psychological theory on her own, this would still involve what we may call third person criteria for application of such states.

To conclude, the myth explains both why it is that we seem to have direct access to our psychological states - because we do to some extent have such access - and why this doesn't lead to solipsism. The consequences for self-knowledge we will see below. The myth captures one way in which we can flesh out the idea that all theory theorists clearly must be committed to: in attributing psychological states to themselves, subjects must in some essential way draw on their knowledge of folk psychological theory.

how?

8. *Theory Theory & Accounts of Self-Knowledge*

In this section, I want to look at some recent accounts of self-knowledge to see whether they are compatible with Theory Theory and, if not, how they are incompatible with it. I will not consider accounts of self-knowledge that are seriously problematic, like behaviourism, expressivism, and inner perception theories. The aim is not to determine which account of self-knowledge is correct or whether a theory theorist is free to choose *any* account of self-knowledge. It is simply to indicate that what Theory Theory has to say about self-knowledge, is by no means fundamentally different from what many philosophers want to say about it anyway.

A good place to start is with functionalist theories of self-knowledge. Whereas there may be functionalists that champion a fully symmetric account of psychological attribution, there are certainly plenty that don't. There are at least two different functionalist accounts of self-knowledge, the 'classical' one and the more modern one (for lack of a better word). According to classical functionalism, as expressed by Brian Loar (1993) and Sydney Shoemaker (1990 & 1993)[45], it is part of the causal role of certain psychological states that, under certain conditions, if you have them, then you believe that you have them. These psychological states are: (Shoemaker, 1990, p. 188)

not in biblio.

[45] Shoemaker (1993) says that classical functionalism is near enough his own view, so minor discrepancies can be expected.

> sensory states, including both sensations (e.g., pains) and perceptual states (e.g., seeming to see red), and intentional states, such as beliefs, desires, and intentions. One claim is that such states are necessarily "self-intimating": that it belongs to their very nature that having them leads to the belief, and knowledge, that one has them, or at any rate that it normally does so under certain circumstances. Another claim is that a person has "special authority' about what such states he or she has.

Specifying (some of?) these conditions, we get the following claim. If you are rational, you have the concept 'belief', you have the belief that *p*, and you consider whether you believe that *p*, you will come to believe that you believe that *p*.[46] This is normally explained with reference to some subpersonal or neural mechanism (Shoemaker, 1993; Loar, 1993). According to Loar, the self-ascriptive process takes you from the belief that *p* to the belief that you believe that *p*. The latter tracks the former, whilst itself being a realiser of a self-ascriptive state with the content that it tracks. Shoemaker stresses that however one wants to conceive of self-ascription, one should not conceive of it as involving any mechanism over and above that which is involved in implementing the belief and the self-ascriptive belief themselves. This is due to the fact that such an implementation must be an implementation of inferential role, which should itself include the self-ascriptive mechanism.

What I have called the more modern functionalist position, has been suggested by, for example, Georges Rey (1993) and Kim Sterelny (1993). According to it, first personal attribution is based on the assessment of *characteristic* rather than defining features of psychological states. Characteristic features are directly accessible to the subjects themselves but not to anyone else. They are reliably connected with the relevant functional roles. So although a psychological state is always defined in terms of the functional role it occupies (or its functional state), self-attributions may be based on evidence other than the presence of some of the relevant causal

(I thought G had a reply to this.

[46] Two points are in place here. Firstly, the specification of the states that are self-intimating and the circumstances under which they are so, are meant to rule out unconscious psycho-analytic and cognitive states (Shoemaker, 1990, p. 188). Secondly, presumably the formation of third-order psychological states on the basis of second-order psychological states, will follow a similar pattern, and so on for fourth-order psychological states, fifth-order psychological states, etc.

factors. This does mean that self-attributions won't be a hundred per cent reliable but, as we have seen, they don't appear to be either.

A challenge that faces all functionalists is to account for the kind of mistakes that we make in self-attributions and the ones that we don't. It is not immediately transparent how this will work out on an account in terms of subpersonal processes. The challenge for the classifaction-due-to-characteristic-features-functionalists is to show how our fallibility is due to classification according to characteristic and not defining features. The impossibility of error through misidentification is not difficult to explain. The fact is that the evidence that we are presented with, albeit only characteristic features of a functional state, are not features that we are presented with when we consider the functional states of other people. It may be more difficult to explain the authority of self-ascriptions. For it seems possible that an intelligent being furbished with a good theory could be more authoritative than you about your psychological states. However these problems may be solved, we see that both these accounts are partly asymmetric accounts of self-attribution, and both are compatible with the idea that we do have some distinctive knowledge of our own psychological states, that we don't have of those of others.

Lastly, let us consider Peacocke's account of self-knowledge. According to him, it is built into the possession conditions of our psychological concepts that if we possess them, then we will be disposed to self-attribute the conscious psychological states that we have.[47] The possession conditions for the concept 'belief' must have at least two clauses,[48] both of which must be known by whoever is to be attributed mastery of that concept (cf. also Strawson, 1959). One clause applies to the first person, present tense case, and the other to the third person case: (Peacocke, 1992, pp. 163-4)

> A relational concept *R* is the concept of belief only if
>
> (F) the thinker finds the first-person content that he stands in *R* to *p* primitively compelling whenever he has the conscious belief that *p*, and he finds it compelling because he has that conscious belief; and
>
> (T) in judging a thought of the third person form *aRp*, the thinker

[47] In *A Study of Concepts*, Peacocke only talks of 'belief'. I take it, however, that he wants to provide an account of self-knowledge in general, not just of knowledge of own beliefs. His formulation there, then, provides the basis of such an account.

[48] Peacocke is not claiming to provide the full possession conditions.

> thereby incurs a commitment to *a*'s being in a state that has the same content-dependent role in making *a* intelligible as the role of his own state of standing in *R* to *p* in making him intelligible, were he to be in that state.

My self-attribution, say, of the belief that *p*, is a case of knowledge that I believe that *p*, because of the combined fact of the possession conditions of 'belief' make it a condition that whenever I (being an instance of a subject) have a conscious belief, say the belief that *p*, I am willing to judge that I have that belief, and the fact that I judge that I believe that *p* for the very reason that I consciously believe that *p* (p. 157).

Peacocke's account is not a million miles away from a classical functionalist account of self-knowledge. Peacocke, however, is opposed to such a view because, according to him, it fails to give *reasons* for one's self-knowledge. It only explains our entitlement to a distinctive kind of self-knowledge in terms of causes, not in terms of reasons. That, to Peacocke is not satisfactory, and he takes his own account to provide the requisite reasons (Peacocke, 1998). Just as Peacocke objects to a classical functionalist account of self-knowledge, so there are certain obstacles for a classical functionalist to embrace Peacocke's account. There are at least two reasons for this. Firstly, there is the issue about reasons and causes. Whereas a functionalist is probably very happy to talk of reasons, as long as it is understood that they are really just psychological causes, it is doubtful that a functionalist would settle for intelligibility in the third person clause, making no reference to causation. Secondly, functionalists are likely to require that there be some mention of behaviour in the possession conditions of a psychological concept, since the production of behaviour is a typical effect of many psychological states, including belief. Peacocke does not mention this at all. It might also be thought that there is a third reason. Peacocke says that "there is a sense in which the concept of belief is a first-person concept" (p. 164). Could a functionalist accept this? In order to determine this we must first see why one might think that functionalism is incompatible with this idea, and secondly, what exactly Peacocke can have in mind.

According to semantic common sense functionalism, the meaning of psychological terms is given by the role that they play in folk psychological theory. It is then natural to expect that for people to have psychological concepts, they must know folk psychological theory. However, if we are really talking about knowledge of a theory, does it make sense to talk of the concepts involved in such knowledge being first personal? That is, isn't there a sort of objectivity involved in something being a theory, such that it makes no sense to require that in order to know that theory, one must be in possession of first personal concepts? Knowing folk psychological theory falls under the category of knowledge of theories more widely. In order that I know, say, probability theory, I need not possess first person concepts. Indeed, it seems to make no sense to require this. If we understand 'theory' on the model of 'scientific theory', then it does seem plausible to claim that theories are relatively objective (cf. Nagel, 1986). However, a theory cannot be relatively objective if knowledge of it involves possessing first person concepts. Another way to look at it is this way. Imagine aliens land and start exploring the world. They learn to communicate with us, learn our customs and ways. They also say things like "Oliver believes that there is more to this than meets the eye" and "Mary Ann is terrified of heights". They also profess that they have no emotions and from what we can tell, they don't. They feel no fear, no joy, no love. Nevertheless, they are perfectly proficient in attributing such states to others - they make the right kinds of reports under the right kinds of circumstances, behave as if they were expecting certain emotional reactions from others and so on. This would be perfectly compatible with functionalism. These aliens would have concepts of emotions. In other words, there would be nothing first personal about these concepts.

Now, does Peacocke mean anything so substantive by 'first person concept' that a functionalist is forced to take issue with him? Let us assume that the two clause possession conditions hold for all psychological states that we are aware of having. (F) and (T) are perfectly compatible with the aliens scenario. If the aliens did have emotions, they would come to self-ascribe such emotions just in the kind of circumstances specified by (F). Furthermore, we can regard the aliens being fully committed to (T) also. It is just the case that they never have emotions, so the states of affairs specified by the two clauses are never actualised.[49] However, Peacocke talks of a capacity

[49] I do not take this to be uncontroversial as an idea. I do take it, however, that it is not obviously false.

to self-ascribe as lying at the core of possessing the concept 'belief'. Extending this to emotion concepts, we might ask what sense it would make to ascribe to the aliens a capacity to self-ascribe emotions according to the conditions outlined by (F)? Is a capacity nothing more than the following counterfactual holding true; if one had emotions and concepts of these emotions, one would be willing to self-ascribe oneself such emotions whenever one was conscious of such emotions? This is decidedly a thin notion of 'capacity', as is made clear by our case of the aliens. I would therefore be inclined to interpret Peacocke as denying that the above alien scenario would be possible or, at least, that these aliens could ever possess the emotion concepts. If this is right, then Peacocke's account is importantly different from a classical functionalist account.

Not surprisingly, either functionalist account of self-knowledge is compatible with Theory Theory. However, semantic functionalism is built into both accounts. As I said above, a theory theorist is not required to accept this position, nor is she required not to do so. If we combine Theory Theory with semantic functionalism, we end up with a very good explanation of why knowledge of folk psychological theory is essentially involved in our self-attributions. This is due to the fact that our psychological terms are defined in terms of this theory. This is the reason that children can fully self-attribute psychological states only once the rudiments of folk psychological theory have been acquired. The fact that in order to acquire the concepts, one must acquire the theory, does require there to be some form of boot strapping procedure, whereby some knowledge of the theory makes one form certain concepts that then allow one to acquire more theory and eventually to restructure or change one's concepts. This, however, seems entirely plausible.

elaborate

Theory Theory without semantic functionalism will have to take a different turn; but not very different. Even if we deny that folk psychological theory is definitive of psychological concepts, we need not deny that there is a very close connection between the two. It seems folly to do so. The knowledge embodied in folk psychological theory clearly contains much of what is involved in psychological concepts. Whereas it is true that were folk psychological theory proved false, we might still turn out to have psychological states, it will also be true that our concepts thereof would be very different. Only, they wouldn't be so different that we would be unable to say that folk psychological theory was wrong about psychological states. So, there is an intimate connection between psychological concepts and folk psychological theory: in order to acquire the former, one must acquire at least rudiments of the latter. This seems like a reasonable position, although it would need to specify more exactly the connection between the concepts and the theory.

As it stands, Theory Theory is not compatible with Peacocke's theory. Most importantly, if Theory Theory takes seriously the idea of 'theory', then it is committed to accepting that the alien scenario is possible, just like functionalism is. I have been arguing all along that Theory Theory *should* take the idea of 'theory' seriously. I, therefore, submit that Theory Theory is committed to the alien scenario being possible and, hence, to denying Peacocke's account - at least on the interpretation that I have given it. Notice, however, that the difference between a Peacockian account and a Theory Theory account is not great. It seems exaggerated to say, as Heal does, that the latter account is resolutely third personal. Commitment to an alien scenario shows that there certainly are certain commitments to the *objectivity* of psychological concept~~s~~ ascription?, but this does not boil down to the fact that we do not have distinctive first personal grounds for both self-attribution and self-knowledge.

9. Constructing a Theory Theory Account of Self-Attribution & Self-Knowledge

We have seen that Theory Theory is compatible with some accounts of self-knowledge, or compatible with minor reworkings of some such accounts. This indicates that the position Theory Theory takes *vis-à-vis* self-attribution and self-knowledge is by no means singular. It is not resolutely third personal in any robust sense. It is less first personal than some accounts self-knowledge, for example Peacocke's'. However, this does not make self-attribution symmetrical, nor does it imply that we are not entitled to a distinctive kind of self-knowledge. We have also seen that even if one opts for a Theory Theory that is also semantic functionalist, one is still not committed to attributional asymmetry nor to the impossibility of us having

Direct access –

1st order ⇒ 2nd order.

This can happen, crudely, without ~~theory~~ FP.

But becomes more sophisticated as FP is acquired. So there is a ~~[illegible]~~ non-TT component in this account of self-K.

distinctive knowledge of our own minds. The same should be true for metaphysical functionalism.

Why isn't this just a retreat from TT?

Self-attribution is most plausibly seen as *sometimes*, at the very least, being grounded on *direct* access to one's own psychological states.[50] On the other hand, such access alone cannot give us self-attribution. Theory Theory holds that we have some direct access to some of our psychological states, but that knowledge of folk psychological theory must play some role in self-attribution also. In the myth of the ontogenesis of self-attribution, we got some idea of how this works. It is part of the nature of certain first-order conscious psychological states - sensations and intentional states - that they tend to give rise to[51] second-order psychological states to the effect that the subject is in the first-order psychological states, *on the condition* that the subject is suitably cognitively, doxastically and conceptually equipped. How exactly one wants to flesh out this idea is up to the individual theory theorist. As I indicated above, whether or not one wants to combine one's Theory Theory with semantic functionalism, acquisition of psychological concepts and folk psychological theory will go hand in hand. The consequences of holding such views will differ, however.

Theory Theory need not hold what Peacocke (1998) calls a no-reasons view. Shoemaker's position is a no-reasons view because he bases self-knowledge on some rote neural mechanism. This mechanism is supposed to explain why it is that when one has a belief that *p*, say, one will come to believe that one has the belief that *p*, provided that one possesses the concept of belief. Alternatively, one might provide an explanation according to which having a belief that *p* is a reason for having a belief that one believes that *p*. For example, it is a combination of the nature of consciously accessible psychological states, themselves, and the nature of folk psychological concepts that, when an organism is in a state of the former kind and possesses the latter, it will come to ascribe itself the psychological state in question.

Let me recapitulate why the above position need not be semantic functionalist. It is obvious, I think, that psychological concepts and

[50] However, again it is worth pointing out that Theory Theory must allow for the possibility of creatures that have no such access, but always self-attribute on third personal grounds. This, however, does not appear to be the case with humans.

[51] I shall not go into how to flesh out 'giving rise to'. Let me just indicate that I don't think it needs to be a causal analysis - i.e. one in which first-order psychological states *cause* second-order psychological states.

folk psychological theory are intimately linked. It is not, however, obvious that the meaning of psychological terms is given solely in terms of the theory in which they figure. Psychological terms can be regarded as rigid designators. Let us assume that what speakers need to know in order to know the meaning of folk psychological terms is that: i. the term is a folk psychological term, ii. the referents of such terms are representational psychological states, and iii. that psychological states have characteristic causal powers to the effect that they are caused by states of the environment and other psychological states, and they cause states of the environment and other psychological states in their turn. This is a bit more complicated than what speakers need to know in order to know the meaning of physical magnitude terms, but it certainly seems to be in the same ballpark. The rest of the meaning of the terms might be given by their reference, as in the case of physical magnitude terms. I do not wish to advocate such an account here, but merely indicate that something like it is open for the theory theorist.

Let me address one last question. If what I have said so far is true, how come it is so frequently thought both that Theory Theory is a functionalist theory and that it is committed to attributional symmetry and denying our entitlement to a distinctive kind of self-knowledge? There is certainly a *historical* link between Theory Theory and functionalism. Sellars was an early functionalist and also the first to propound the Theory Theory position. His aim was to illustrate how we might come to think of ourselves as subjects of psychological states, without the psychological states themselves being given to us directly (Sellars, 1963). It is important to note, however, that Sellars did not seem to embrace attributional symmetry. He maintained that we can be trained to "give reasonably reliable self-descriptions ... without having to observe [our] overt behaviour" (p. 189) and that, although there is no absolute privacy of subjects *vis-à-vis* their "inner episodes", there is nevertheless some form of privacy. Sellars' proposal proved remarkably influential. Dennett, for example, traces the idea of folk psychological theory back to him (Dennett, 1987). When Lewis (1972) takes up the Theory Theory idea, it is in the context of arguing for semantic and metaphysical common sense functionalism. Presumably all this is where the idea that Theory Theory is a functionalist idea stems from.

There is, however, another origin of the idea. Child psychologists use the term 'theory of mind' largely in the same way that philosophers use 'folk psychology'. However, originally both the phrases seemed to denote Theory Theory. The term 'theory of mind' was introduced by Premack & Woodruff (1978). According to them, someone has a theory of mind if he: (p. 515)

> imputes mental states to himself and to others (either to conspecifics or to other species as well). A system of inferences of this kind is properly viewed as a theory, first, because such states are not directly observable, and second, because the system can be used to make predictions, specifically about the behavior of other organisms.

Someone like Fodor, who is clearly a theory theorist, simply talks of folk psychology (1987). Later, with divergence in opinion concerning how to explain our practice of attributing psychological states to one another, the term Theory Theory was introduced (Morton, 1980). Now, Theory of Mind is used in psychology to denote the practice of attributing psychological states to one another, not necessarily to refer to Theory Theory. Nevertheless, there is little doubt that when psychologists talk of Theory of Mind and Theory Theory, they mean to latch on to the experimental tradition originating in Premack & Woodruff's work, not Sellars' early functionalism.

So much for the history of the mistake. Theory Theory is now a term that refers to an internal account of folk psychology that is largely empirical in character. Common sense functionalism is a somewhat different story. It need not be an internal account, among other things. Theory Theory and functionalism can be combined, but need not be so. Where the idea came from that functionalists are committed to a symmetric account of psychological attribution and must deny that we have knowledge of our own psychological states, I'm not sure. As I have pointed out, it is not even to be found in Sellars.

Chapter 5

Theory Theory & Tacit Knowledge

We have a very extensive shared understanding of how we work mentally. Think of it as a theory: FOLK PSYCHOLOGY. It is common knowledge among us; but it is tacit, as our grammatical knowledge is. We can tell which particular predictions and explanations conform to its principles, but we cannot expound those principles systematically. (Lewis, 1994, p. 416)

Among the many cognitive capacities that people manifest, there is one cluster that holds a particular fascination for philosophers. Included in this cluster is the ability to *describe* people and their behavior (including their linguistic behavior) *in intentional terms* - or to 'interpret' them, as philosophers sometimes say. [...] Since the dominant strategy for explaining any cognitive capacity is to posit an internally represented theory, it is not surprising that in this area, too, it is generally assumed that a theory is being invoked [...] The term 'folk psychology' has been widely used as a label for the largely tacit psychological theory that underlies these abilities. (Stich & Nichols, 1995a, pp. 123-4)

All along, I have been talking about our knowledge of folk psychological theory. We have already seen what this knowledge amounts to in terms of what the *content* of it is. It is now time to turn to its *form*. Forms of knowledge have been much discussed in recent decades; in particular, tacit knowledge versus ordinary knowledge (Chomsky, 1975; Davies 1989b; Fodor, 1981; Marr, 1982; Stich, 1978). A number of theory theorists claim that knowledge of folk psychological theory is tacit (Braddon-Mitchell & Jackson, 1996; Lewis, 1994; Stich & Nichols, 1992). In this chapter, I will examine what this claim amounts to and whether it should be accepted as forming part of Theory Theory.

I shall proceed as follows. First, I will look at transformational grammar and how tacit knowledge figures in this area. Tacit knowledge was first introduced in linguistics and vision research, and to see what role it is supposed to play in folk psychology, it is important to understand how it figured originally. How tacit knowledge is related to unconscious knowledge will be addressed here and clarified later. Second, I shall examine three prevalent philosophical theories about the nature of tacit knowledge. We can then move on to see whether our knowledge of folk psychological theory has any of the features of tacit knowledge. This is the third step. I rely on a comparison between knowledge of grammar and knowledge of folk psychological theory. For this comparison, we need specific examples of the statements of folk psychological theory. And since tacit theory theorists do not tend to provide any examples of folk psychological generalisations, I rely on the kind of generalisations given in chapter 1. We shall see that none of these generalisations appear to be tacitly known. Fourth, I look over some of the recent research in experimental psychology concerning what psychologists call 'implicit learning' and 'implicit knowledge', but which is synonymous with tacit learning and knowledge. The aim is to discover whether we here find a radically different view of tacit knowledge that might endanger my conclusion that folk psychological knowledge is not tacit. We don't. According to the only radically different suggestion - that tacit knowledge isn't representational - folk psychological knowledge will still be explicit.

1. Tacit Knowledge of Transformational Grammar

'Tacit knowledge' is a technical term used primarily in cognitive psychology and linguistics. It denotes a collection of cognitive states that are assumed to be explanatory of a particular ability that a subject has, such as the ability to utter and comprehend grammatical sentences. To use 'knowledge' here, albeit prefixed by 'tacit', is, perhaps, infelicitous. The term, in its standard philosophical use, connotes truth and justification. *Tacit* knowledge, however, has no commitments to either. It simply refers to a body of representational states that is causally efficacious in the production of a specified range of behaviour. Such states are also known as subdoxastic states. This choice of vocabulary is more apposite, 'doxa' being Greek for 'belief' or 'opinion'. Often, subdoxastic states are contrasted with beliefs, but sometimes also with the entire class of propositional attitude states. Usage of the term 'subdoxastic' stresses the fact that the contrast between tacit knowledge and ordinary knowledge was never supposed to be between different kinds of knowledge as such, but between different kinds of cognitive states.

Knowledge of grammar - alongside knowledge of visual parameters (Marr, 1982) - is the prototypical example of tacit knowledge. The idea is intimately linked with the figure of Chomsky (e.g. Chomsky, 1975). According to him, tacit knowledge of grammar enters into the explanation of our linguistic ability, more precisely the ability to utter and comprehend grammatical sentences. Transformational grammar is the particular grammar that we are assumed to have tacit knowledge of. In general, it looks very different from the kind of grammar one is familiar with from early schooling.

Transformational grammar works with two or more levels of representation.[52] I shall give an example of the principles of

[52] How many levels are posited depends on which phase in Chomsky's work one looks at. He started out with two, at a point had four, and now works with no more than three levels. Fortunately, the exact number of levels is not important to the argument at hand.

transformational grammar in terms of a three-level structure. This structure is composed of the following levels: DS, LF, and PF. These are technical terms used by Chomsky connoting, respectively, deep structure, logical form, and phonological form. The DS of a sentence is the basic grammatical structure which is constructed out of the lexicon and basic grammatical rules, such as those of X-bar Theory. The PF is identical to the heard sentence; that is, it is the level where the characteristic sounds of language are represented. LF is where anaphora, scope, and the like are represented. According to Larson & Segal (1995), it is here that syntax interfaces with semantics. At this level, the DS of a sentence has been transformed. Rules, other than those operating in the DS, govern these transformations.

As a concrete example of tacit knowledge of grammar, let us look at *wh*-traces. A *wh*-trace is a trace left in the LF of a phrase where the interrogative pronoun - *what, who, which, when*, or *where* - has moved from the position it occupies in the DS of the relevant phrase. The trace is left in the LF where the *wh*-word figures in the DS such as to capture the scope of the verb of which it is an argument (see fig. 1). Now, 'like' is a transitive verb and, as such, it takes both a subject and an object. At the level of DS, the subject will appear to the left of the (transitive) verb and the object on the right, as is apparent in "John likes whom?". At the level of PF, the interrogative pronoun has moved relative to the position it had in the DS. The *wh*-trace in the LF indicates the position of the relevant interrogative pronoun in the DS.

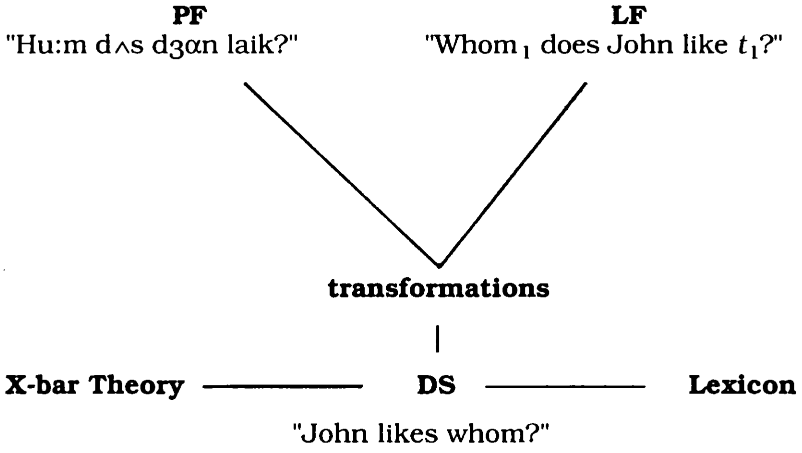


Fig. 1

In fig. 1, the fact that the trace refers to 'whom' is represented by the '$_1$' tagged at the end of both trace and interrogative pronoun. This safeguards our understanding the sentence as one in which 'whom' is the object of 'like' (cf. Chomsky, 1986, pp. 77-78).

According to Chomsky, what best explains our speech behaviour is knowledge of grammatical rules such as those pertaining to *wh*-traces - along with certain performance mechanisms that access this knowledge and put it to use. I cannot go into the arguments here, but suffice it to say that this is the picture that forms the basis of most theorising about the nature of tacit knowledge. In what follows, I shall use our knowledge of *wh*-traces as representative of our grammatical knowledge. What is true of this knowledge, I will assume is true of all of grammatical knowledge.

Tacit knowledge works like ordinary knowledge in a number of respects. It is, for example, causally efficacious in the production of a certain range of behaviour. In other respects, however, it is unlike ordinary knowledge. Most of us are unaware of possessing it, for example. There is real dispute concerning what characteristics set tacit knowledge apart from ordinary knowledge, and what consequences that has for psychology. A corollary of this is that what is classified as tacit knowledge, over and above grammatical competence and vision, is not generally agreed upon.

2. Chomsky & Stich: Tacit Knowledge Is Unconscious

Noam Chomsky has suggested that the only difference between subdoxastic states and beliefs is that subjects attributed these states are aware of having the latter, but unaware of having the former. According to his view, the two sorts of states are both cognitive states of the following kind. They are representational, causally efficacious in the production of thought and behaviour, and their causal powers are associated with their representational contents. Unconscious psychological states *qua* unconscious are not likely to constitute a psychologically interesting class. Hence, this feature of subdoxastic states is irrelevant to psychological theorising. We, in our role as cognitive psychologists, should free ourselves of our pre-scientific psychological concepts, like 'knowing' and 'believing', since the significant notion in future psychology will probably be something like 'cognizing' (Chomsky, 1976).

It is not entirely clear what Chomsky has in mind by 'unconscious' here. If we turn to Stich's theory of tacit knowledge, we will see how it is possible to mean at least two different things by 'conscious', and consequently the negation of it. Stich (1978) thinks that the difference between subdoxastic states and beliefs is far from being unimportant to cognitive psychology. Subdoxastic states form an importantly different class of cognitive states from beliefs, because they are consciously inaccessible and inferentially encapsulated, whereas beliefs are neither.

A belief is conscious, according to Stich, if we have a characteristic conscious experience when our attention is suitably drawn to the content of it. Generally, a subject will assent to a proposition that she believes to be true. However, conscious experience is only connected with verbal assent insofar as the subject is *disposed* to assent to a proposition expressing the content of one of her beliefs. Assent and conscious experience are interconnected, but the former is not essentially correlated with the latter. Rather, it is "the experience of having an occurrent belief" that is at issue (p. 504).

One imagines that there must be many varieties of characteristic conscious experiences; perhaps as many as there are psychological modes. Fear, pain, anger, and passion are all high on phenomenological impact. Beliefs, on the other hand, presumably have less phenomenological import. Indeed, it is hard to imagine what phenomenology belief could have other than something like a gut-feeling, a feeling of recognition, or of unsurprise. To take an example, when you are presented with the content of a belief that you have, say:

(1) Paris is the capital of France

you should have some kind of feeling of recognition or unsurprise. This may not be terribly salient, but should be salient enough for you to be able to distinguish it from the phenomenology of being presented with the content of a belief that you don't have, say:

(2) Nauplion was the first capital of modern Greece

In this case, you should have a feeling of mild surprise, informativeness, or something of that sort. It is, at any rate, a feeling qualitatively different from that which (1) elicits. Hence, the phenomenology connected with being presented with contents of beliefs that one has and contents of beliefs that one does not have is recognisably different. Only in the former case do we have a characteristic conscious experience.

Beliefs have roughly the same phenomenology whether or not they are explicit or implicit. Take:

(3) Cars aren't living organisms

This is a traditional example of an implicit belief - a belief that you have never consciously entertained, but that is implied by other beliefs that you *have* consciously entertained. However, although you may never have thought of this before, it is not like (2) - it is neither surprising, nor unfamiliar.

Subdoxastic states are also inferentially impoverished or encapsulated. The inferential patterns by way of which they can give rise to beliefs and beliefs can give rise to them, are extremely limited. Beliefs are well integrated into an inferential network of other beliefs. For instance, if I believe that Ted Honderich is the author of *Violence for Equality* and that the author of *Violence for Equality* was the Grote Professor at UCL, I can infer that Ted Honderich was the Grote Professor at UCL. Subdoxastic states might form an inferential web with certain other subdoxastic states but not with others.[53] Subdoxastic states whose contents concern *wh*-traces will not play a role in any inference, involving states whose contents concern visual parameters. Beliefs stand in causal relations that subdoxastic states don't and *vice versa*. Notice that this does not mean that beliefs and subdoxastic states are not related to each other inferentially at all. Subdoxastic states regularly give rise to beliefs - the operation of our tacit knowledge of grammar gives rise to the belief that someone is saying this or that. However, the range of beliefs that a particular subdoxastic state can give rise to is importantly restricted compared to

[53] Stich does not actually think that subdoxastic states can form part of an inference, only beliefs can do so. When he talks of inference involving subdoxastic states, he means to talk of inference-like psychological operations (1978, pp. 511-17).

the range of beliefs a belief can give rise to. Furthermore, it is doubtful whether beliefs and subdoxastic states together can inferentially give rise to other beliefs or subdoxastic states.

The notion of inferential encapsulation is best fleshed out in terms of cognitive subsystems (Stich, 1978; Davies, 1989). Beliefs "form a consciously accessible, inferentially integrated cognitive subsystem" whereas subdoxastic states "occur in a variety of separate, special purpose cognitive subsystems" (Davies, 1989, pp. 507-8). We can put the idea this way:

> A cognitive subsystem is **either**:
> i. a body of psychological states that are interrelated in terms of their component concepts, **or**
> ii. a special purpose processor, **and**
> iii. encapsulated from information outside it

What makes us want to talk about a cognitive *sub*system is that there is an identifiable part of the cognitive system - either in terms of function or information - that is separable from the functioning of other subsystems and the system as a whole. A subsystem is idiosyncratic either in terms of its processes or its information or both, and is only sensitive to certain kinds of information.[54] It is the latter that constitutes its informational encapsulation.

Philosophers and cognitive psychologists often talk of beliefs forming a cognitive subsystem: the belief box. The belief system stores only beliefs. All the beliefs are sensitive to each other. Beliefs are not sensitive to subdoxastic information until it is presented in a belief format. There doesn't seem to be a corresponding subdoxastic box. That is, subdoxastic states are stored in a number of separate boxes - one or more of which concern language and store grammatical rules, one which concerns vision and stores visual parameters, and so on. For this reason, subdoxastic states cannot interact with each other as beliefs can among one another. Furthermore, the encapsulation is

[54] Strictly speaking, this is incorrect. A subsystem is sensitive both to what information is presented to it and the provenance of it. Thus, even if the information concerns the right subject matter, it can still fail to penetrate into the relevant subsystem. For example, the visual subsystem is not sensitive to information about visual processing that a subject may acquire studying optics. The visual system is only sensitive to information received from the retinas.

such as to make only a limited amount of information available to other cognitive subsystems.

On a more encompassing picture of the cognitive system, the informational encapsulation of subsystems alone cannot explain inferential encapsulation. Once you begin to populate the cognitive system with more subsystems, the need arises for a place of interaction between at least some of those subsystems. Systems like the belief system and the desire system - assuming that there is a desire system corresponding to the belief system - must be more or less integrated with each other, such that the information in both can come together somewhere to form intentions, prompt actions, and so on. It must be possible for states contained in at least some of the subsystems to interact. One way to safeguard such interaction, is to assume that subdoxastic subsystems have filters that prevent most, or all, of the information contained in them to be broadcast outside the system. Thus, the free inferential interaction of propositional attitude states with each other is due to the fact that the subsystems that they form part of do not have such filters. Of course, such a filter should not exclude there being privileged interaction between the information of some subdoxastic subsystems. For example, we need the grammar system and the lexical system to interact in language production and comprehension. Another way to secure the interaction between propositional attitude states is to assume that there is another cognitive subsystem, a sort of cognitive workspace (cf. Bernard Baars' global workspace theory of consciousness (Baars, 1988)) that is sensitive to the informational characteristics of propositional attitude states but not to those of subdoxastic states. Figure 2 gives us a flavour of how such a cognitive system would look.

Both conscious accessibility and inferential integration can be regarded as conscious features. Having some information generally available in one's theoretical or practical reasoning is one aspect of what it is for information to be conscious. The other, is the phenomenological aspect we talked about above, where there is a sense of recognition when one is presented with the content of such information. I don't think Chomsky is easily interpreted as holding meaning 'inferentially encapsulated' by 'unconscious', since he would then commit himself to holding that such a feature is irrelevant to cognitive psychology. That hardly seems plausible. Therefore, we

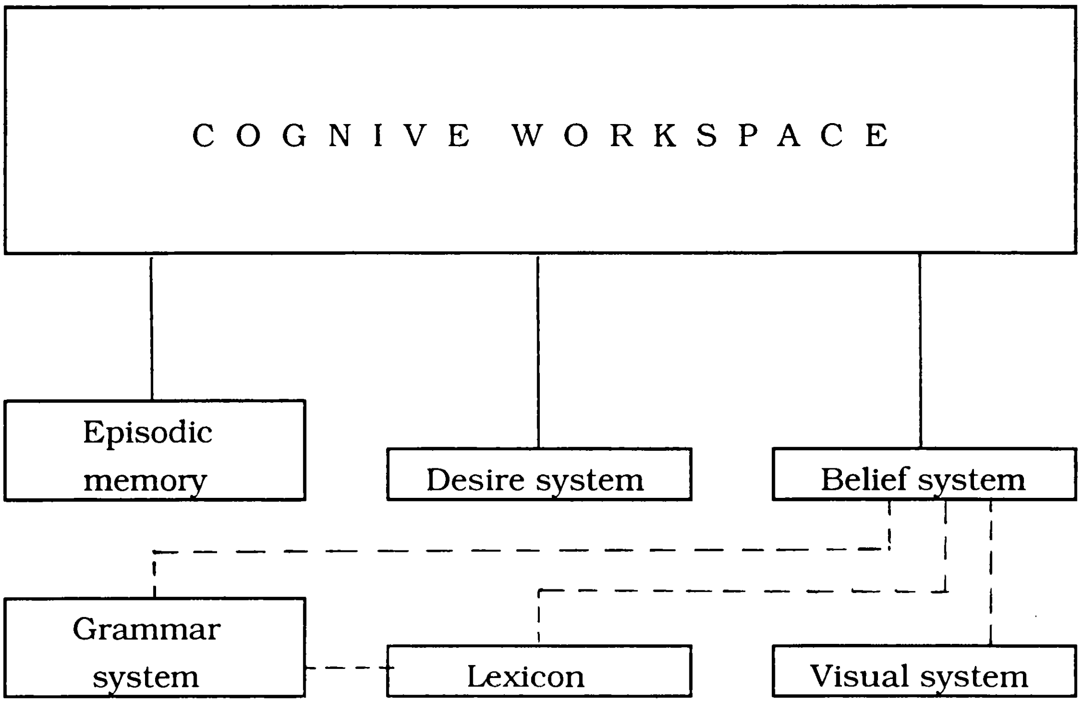


Fig. 2

should assume that Chomsky means 'consciously inaccessible' by 'unconscious'. So, there is a real disagreement between Stich and Chomsky about what features tacit knowledge has. This may be what is at the root of the disagreement about what consequences assuming that we have tacit knowledge has for cognitive psychology.

Let us now return to our knowledge of *wh*-traces to consider whether it has any of the above features. I use this as a test of any theory of tacit knowledge that it accords with knowledge of such traces. Firstly, we don't seem to be aware of our knowledge of *wh*-traces. When looking at figure 1 and the surrounding description, I take it that no reader untrained in linguistics had a characteristic conscious experience. Secondly, our knowledge of *wh*-traces is not inferentially integrated with our beliefs, desires, emotions, and so on. Our control over this knowledge is strictly limited to the utterance and parsing of grammatical sentences. We cannot, for example, sit down and reflect on this knowledge to work out an internal account of grammar. We have to look at the judgements that we make, not introspect on our pre-existing knowledge. Of course, grammatical knowledge is not completely inferentially encapsulated, it interacts with other parts of our knowledge of language - for example our lexicon. Nevertheless, grammar is applied to language production and comprehension exclusively. Consequently, it appears that our grammatical knowledge is consciously inaccessible and inferentially encapsulated.

3. Davies: The Generality Constraint

Davies has suggested the Generality Constraint as providing a principled distinction between beliefs and subdoxastic states.[55] The Generality Constraint is a constraint on thought first suggested by Gareth Evans (1982). It is applicable to the issue at hand in the following way. Most people accept that thoughts are structured states. They are composed of a number of concepts. As a rule, we don't attribute to somebody the thoughts 'the cat is on the mat' and 'the girl is in the garden' if they cannot also entertain the following two thoughts: 'the cat is in the garden' and 'the girl is on the mat'. According to the Generality Constraint:

> for a subject to have the thought *Fa*, she must be able to: i. conceive of *a being G, H, J*, and so on for all properties she knows of, and ii. conceive of *b, c, d*, and so on *being F* for all the objects she knows of.[56]

This constraint is only formulated fully for propositions of the subject-predicate form, but we must imagine that it applies to all possible propositional forms (Davies, 1989). The constraint holds on thoughts, not on propositions. It concerns the relation between a subject and a proposition. It stipulates what kind of dispositional state a subject must be in for her to stand in a relation to a proposition, such as

[55] The distinction only operates on representational states. However, some subdoxastic states are best understood as states of processors, not representational states. Therefore, not all subdoxastic states will be accurately characterised by this distinction. Nevertheless, since all knowledge states are representational, it will suffice to distinguish representational subdoxastic states from belief states. It may eventually be complemented. For example, because all beliefs are representational, beliefs differ from subdoxastic states on two dimensions: i. beliefs are representational, some subdoxastic states are not, ii. beliefs differ from representational subdoxastic states by being subject to the Generality Constraint.

[56] This is a condensation of Evans (1982), pp. 103-4.

believing, desiring, and so on. Subjects lacking such dispositions cannot stand in the relevant relation to such a proposition.

Davies uses this constraint to draw a distinction between subdoxastic states and beliefs in the following way. Beliefs are subject to the Generality Constraint. Assuming that Jane has the concepts 'twelve-tone music', 'baroque music', 'being truly dreadful', and 'stimulating', she cannot be attributed the belief that twelve-tone music is truly dreadful, unless she can also conceive of twelve-tone music being stimulating and baroque music being truly dreadful. Subdoxastic states, on the other hand do not appear subject to the Generality Constraint. Subjects untrained in linguistics are attributed knowledge of the fact that a sentence commencing with an interrogative pronoun will contain a *wh*-trace where the pronoun would figure at the DS. Such an attribution is completely independent of whether the subject can conceive of sentences that begin, say, with a proper name containing a *wh*-trace, even on the assumption that the subject has the concept of 'a sentence that begins with a proper name'.

It is important, Davies warns, not to misunderstand the scope of the constraint. If a psychological state is a belief, it *ipso facto* meets the Generality Constraint. A subdoxastic state might happen to meet the Generality Constraint, but this does not yet show that it is a conceptual state because it is not required *ipso facto* to meet it. What Davies wants to rule out, is the possibility of errant subdoxastic states meeting the Generality Constraint falsifying his idea. This is not an *ad hoc* move, as we are about to see. To illustrate what I take to be Davies' idea, I want to look at the case of a linguist. Linguists along with all other competent speakers can be attributed knowledge of grammar, for example knowledge of *wh*-traces. However, in the linguist's case it appears that her knowledge of these traces meets the Generality Constraint, because '*wh*-trace' can combine with all her other concepts. This points her knowledge being conceptual, and contrast with the situation that holds for subjects untrained in linguistics. The question is, should we say that, in the case of linguists, their grammatical knowledge - that is, the knowledge that all competent speakers possess - is conceptual? It seems to meet the Generality Constraint, and if meeting the Generality Constraint is being conceptual, then this knowledge certainly is conceptual.

Another possibility is to assume that linguists have two separate bodies of knowledge; one that meets the Generality Constraint and one that doesn't. What meets the Generality Constraint is not the knowledge that is explanatory of people's ability to produce and comprehend grammatical sentences, but the knowledge that is explanatory of the linguist's ability to teach, research, and in general to talk about grammar. This idea might seem intuitively implausible, as it appears gratuitously complicated. Nevertheless, I think that there are excellent reasons for accepting it. Firstly, linguists tacitly knew transformational grammar before becoming linguists. For quite a long period, they were in exactly the same position *vis-à-vis* grammar as everybody else. It is a substantial claim to make that simply by acquiring knowledge of linguistics, they restructure the knowledge they already possess tacitly, such as to make it explicit. For all we know this is not possible. Secondly, it seems perfectly possible to imagine cases where the two kinds of knowledge come apart. Imagine a linguist that has been in a serious car crash. She has lost her explicit knowledge of grammar. Whereas she can fully comprehend and produce grammatical sentences, she is unable to explain the principles underlying this capacity. On the other hand, we may imagine that she has lost her tacit knowledge of grammar. In this case, she can explain the principles of grammar, but either has great difficulty uttering or comprehending sentences, or is incapable of doing so at all. It remains an open question whether explicit knowledge of grammar could be effective in speech comprehension and production given the time it is likely to take accessing the relevant parts of this knowledge. Perhaps through practice, the linguist will eventually make this knowledge play the role that the tacit knowledge originally did. However this may be, it is still the case that the two forms of knowledge can come apart in certain situations. The fact of the linguist's early competence and the possibility of her capacity as a speaker and as a linguist coming apart, amount to heavy reasons in favour of accepting the more cumbersome view suggested.

Seeing that a piece of knowledge meets the Generality Constraint, is not sufficient to show that it is explicit *and* causally efficacious in the production of the range of behaviour, it is posited as explaining. When I attribute to the linguist knowledge of *wh*-traces because I want to explain how she manages to utter and comprehend grammatical sentences, I need not assume that this knowledge is conceptual because, although the linguist will act in ways that make me think that her knowledge meets the Generality Constraint, what meets the Generality Constraint is not *this* knowledge, but some other, explicit knowledge that she has, that has the same content. The way to determine whether or not some knowledge is conceptual, is first to test whether it meets the Generality Constraint. If it does, we must make sure that the knowledge that has hereby been revealed as being explicit, is the knowledge that is causally efficacious in the production of the behaviour that it is posited as being explanatory of. One way of doing this is to consider whether something like the linguist scenario holds. Is it reasonable to suppose that the subject possessed knowledge of the relevant information prior to her being able to act in ways that make us want to attribute to her ordinary knowledge thereof? If it is, we can conclude that the knowledge is non-conceptual and, consequently, tacit. If it isn't, it is conceptual and hence it is ordinary or explicit knowledge. In the linguist case, the prior possession of the capacity that is explained by attributing to her the relevant knowledge, as well as the later explicit learning period, is sufficient for us to assume that what is causally efficacious in her production and comprehension of grammatical sentences, is not what meets the Generality Constraint. Consequently, the knowledge is tacit.

A counterexample to this claim comes from knowledge of unconscious psycho-analytic states. People that enter psycho-analysis apparently come to be conscious of some of their (former) unconscious states; they come to have explicit and conceptual knowledge of them. We would want to attribute these unconscious states to them prior to analysis, because they were explanatory of certain, perhaps peculiar, behaviours on their part. Now, the fact that they come to have knowledge that is conceptual surely won't make us say that now they have two forms of knowledge: tacit and explicit knowledge with the same content. Rather, we want to say that the knowledge that they had was conceptual all along, but moved from an unconscious state into a conscious one. But it seems that my above construal of such development disallows this.

It seems to me that there is an important difference between the psycho-analytic patient's case and the linguist's case. Psycho-analytic states come in two varieties: states that are unconscious because they have been repressed, and states that are innate and unconscious (Freud, 1915/1957). It is a reasonable conjecture that whereas the former are conceptual, the latter are not. This allows us to put down the psycho-analytic patient's case to the work of a repression mechanism. This, then, will distinguish it from the linguist's case, where there is no such mechanism in play. Hence, when we are presented with knowledge that we are uncertain about how to classify, we must also consider whether there are good reasons to suppose the subjects to possess the concepts in terms of which we would phrase our knowledge attributions to them, prior to the point where they clearly had conceptual knowledge of the relevant subject matter. If there are such reasons, we should conclude that what we are concerned with are repressed unconscious states having become conscious. In practice, this might not be difficult - the kinds of experiences that make unconscious psycho-analytic states conscious, are normally quite different from the kinds of experiences that are involved in acquiring explicit knowledge about a subject matter that is tacitly known. One might say that both situations might involve learning, but the former is learning and reflecting about oneself, the latter concerns learning something not directly about oneself. In short, I don't think that the psycho-analytic patient scenario poses a serious threat to the above suggestion about how to characterise a linguist's knowledge understood as that which is causally efficacious in the production of her linguistic utterances and comprehension. To conclude, tacit knowledge is not structured by concepts. Knowledge of *wh*-traces accords well with Davies' model. Such knowledge is plausibly seen as failing to meet the Generality Constraint, and hence being non-conceptual.

Lastly, let me point out that the above does not imply that subdoxastic states are not structured - Davies believes that they are. The point is rather that subdoxastic states are not structured by concepts. Nevertheless, the building blocks of such states must be like concepts in some respects. They must map onto some causal role in a cognitive system, and they must be able to combine with a number of other such elements. What distinguishes these quasi-concepts from real concepts is that they do not combine in the fashion required by the Generality Constraint.

4. Folk Psychological Theory

It is not clear that what theory theorists have in mind when they claim that we have tacit knowledge of folk psychological theory, is fundamentally the kind of folk psychological theory that I presented in chapter 1. The problem, however, is that no outline of an alternative folk psychological theory has been presented. It seems to me that we cannot determine whether folk psychological theory is tacitly known or not, unless we know what theory we are talking about - that is, roughly what content it has. I will assume that what these theory theorists have in mind is what I have presented so far as folk psychological theory. And what I have presented so far is consonant with the kind of examples that are presented in the literature. We need not have before us a full formulation of a folk psychological theory in order to be able to ascertain whether it is tacitly known or not. But we need at least a handful of examples to guide us on the way. Therefore, I suggest that we think back to some of the examples of folk psychological generalisations presented in chapter 1. Lets take (G1)* as a typical example to match that of *wh*-traces:

> (G1)* **If** (1) *X* wants to Ø, and
> (2) *X* believes that *A*-ing is a way for him to bring about Ø under those circumstances, and
> (3) there is no action believed by *X* to be a way for him to bring about Ø, under the circumstances, which *X* judges to be as preferable to him as, or more preferable to him than, *A*-ing, and
> (4) *X* has no other want (or set of them) which, under the circumstances, overrides his want Ø, and
> (5) *X* knows how to *A*, and
> (6) *X* is able to *A*, and
> (7) *X* does not believe that the outcome of *A*-ing is such as to make it impossible or too difficult to bring about *Æ*, which is something else that *X* wants as much as, or more than, Ø,
> **then** (7) *X A*-s

It might be objected that although there exists no alternative to this formulation of Theory Theory at the moment, there are two obvious candidates. Baron-Cohen's (1995) theory of the precursors of

ToMM, discussed in chapter 1, and a neo-Chomskian alternative. It might be thought that Baron-Cohen has shown that part of our folk psychological knowledge embodies the information that is contained in the ID, EDD, and SAM. Could the information contained in these modules not constitute the rudiments of an alternative folk psychological theory? Apart from that fact that this view does not appear to be how Baron-Cohen, himself, conceives of his work, it is not really consonant with it either. Understood synchronically, the ID, EDD, and SAM do not *replace* folk psychological theory. Rather, they supplement it in the sense that they make possible the proper operation of this body of knowledge. ID, EDD, and SAM are not themselves theories, they are mechanisms that help you apply the theory (they provide input for ToMM). It is not directly relevant to folk psychological *theory* how one manages to track a direction of gaze, how one detects self-propelled motion, and how one determines that one is attending to the same state of affairs as another individual. The *significance* of all these activities, though, is a part of folk psychological theory. A module like the ID allows you to see certain movements as intentional behaviour. But what allows you to see those movements as something significant is your ToMM - your folk psychological theory. Ditto for the other modules. On the other hand, methods of application are highly relevant to individuals possessing a theory, for to have use of it, they need to know how to apply it. Nevertheless, it is rather unlikely that we can look to ID, EDD, and SAM for alternative formulations of folk psychological theory.

There is another position suggested by the writings of philosophers impressed by the alleged similarity between knowledge of grammar and knowledge of folk psychological theory. As we have seen, the rules of transformational grammar that we tacitly know need be nothing like the ones that our teachers try to imprint on us the first painful years of schooling. Indeed, leafing through a transformational grammar textbook, one is taken aback with the complexity and the unfamiliarity of the rules there presented. Few familiar concepts remain like 'verb', 'adjective', 'subject', and so on, but there is a myriad of concepts undreamt of by linguistic neophytes. Could it not be the case that the representations involved in our tacit knowledge stand to the representations that we are aware of having like the concepts and rules of transformational grammar stand to the grammatical concepts and knowledge that we are aware of possessing?

The view is a kind of extension of the framework of tacit knowledge of grammar to tacit knowledge of folk psychological theory - a sort of neo-Chomskianism. There is one crucial difference, however. Quite a lot of work has been done on transformational grammar, and since subjects don't profess knowledge of these rules, it has been concluded that such knowledge is tacit. There is no comparable situation in philosophy or psychology. Here there is no theory - not even a handful of generalisations suggestive of one. In short, there is nothing to allow us to evaluate the truth of the claim that our folk psychological knowledge is tacit. So whereas the idea certainly merits research, it does not provide us with an alternative of folk psychological theory at present.

5. Folk Psychological Theory and Tacit knowledge

We can now examine whether folk psychological theory is tacitly known. I shall assume that if folk psychological knowledge is consciously inaccessible, inferentially encapsulated, and is non-conceptual, we have good reasons to believe that it is tacit. If, on the other hand, we are to find that it has none of these characteristics, we should conclude that it is not.

Do people have a characteristic conscious experience when they are presented with examples of folk psychological generalisations? To my knowledge there are no experiments to show this. However, first of all, we can turn our attention to (G1) and consider whether that gives us a characteristic conscious experience. As far as I understand Stich, it certainly gives me one. There is another, more indirect way of ascertaining whether folk psychological generalisations are consciously accessible, that relies less on what may seem to be a dubious phenomenological argument. This is connected with verbal report and assent. The idea is that although it is not necessary for a subject to be able to report on her psychological states in order that they be classified as conscious, it seems reasonable to suppose that if subjects report being in a particular psychological state, then they are conscious of being in that state. Likewise, if a subject assents to a question to the effect that she has a certain belief, say, then that belief is consciously accessible.

Of course, not any case of assent indicates conscious accessibility. People lie, assent to propositions that they believe are false, or don't believe are true, when under pressure, and so on. But the fact that there are exceptions does not detract from the fact that under normal circumstances, it is extremely unlikely that subjects are not conscious of the content of what they report or assent to. What we need to ensure, of course, is that any given case is not an unusual one. For example, once we have people's assent to a statement of folk psychological theory, we may want to ensure that the assent is not based on suddenly realising something, by asking apposite questions to that effect. If subjects assent, and deny that their assent is based on a sudden and new realisation, we can reasonably conclude that they are in consciously accessible states that have the same content as the propositions assented to. In practice, we cannot test all the generalisations of folk psychological theory, we will have to limit ourselves to a handful of them. In the absence of the possibility of carrying out such an experiment, we can ask ourselves, do I assent to generalisations such as (G1), and do I have reason to believe that such assent is based on something other than the conscious accessibility of a psychological state with that content? I think you will agree that the answer is: I assent and there is no reason to think that this assent is not based on the conscious accessibility of a psychological state with the requisite content.

In respect of reporting, there are psychological experiments concerning children's understanding of folk psychology, that involve children justifying their psychological judgements. For example, children will explain why they think that a person who is denied visual access to the introduction or displacement of things in a particular location, doesn't know what is there by saying that not seeing implies not knowing (Wellman, 1990; Wimmer, Hogrefe & Sodian, 1988). This, or something very like it, is a folk psychological generalisation when appropriately hedged with *ceteris paribus* clauses. Adults, too, will sometimes discuss folk psychological principles like 'people generally believe what they are told' or 'people like to be flattered'. All this provides extra evidence for the idea that folk psychological knowledge is consciously accessible. Hence, I believe that we have convincing evidence to show that folk psychological knowledge is not consciously inaccessible.

Showing that folk psychological knowledge is consciously accessible by showing that it elicits assent and report under certain circumstances, implies that it is also inferentially integrated. If I can report on some psychological state that I have, not only does it have to be informationally integrated in order that I can do so, but being able to report on it has many consequences for what use I can put that information to. The limits here seem to match those of ordinary knowledge. It is plausible that it is not the fact that I can report on such a psychological state that makes it play the unlimited inferential role in question, but rather that part of what it is to be reportable is already to be able to play such an unlimited inferential role. In addition, beliefs appear freely to give rise to new folk psychological generalisations, and folk psychological generalisations can freely interact with beliefs. All this, taken together, presents a very strong case for the inferential integration of folk psychological knowledge.

Beliefs give rise to folk psychological generalisations in the following ways. Knowledge that we gain from such subjects as experimental psychology, cognitive psychology, and psycho-analysis, profoundly affects the way we think of ourselves and other people. The last couple of decades central ideas of psycho-analysis, such as the idea of the unconscious, have become widely accepted. Now, it will come as no surprise if your greengrocer explains her own or other people's actions by reference to unconscious beliefs or desires. Unconscious motivations and ideas have become part of folk psychology. Likewise, those of us who are well-read in psychology will have updated or changed a number of beliefs about why people think and act as they do, with consequent changes in our psychological attributions. For example, people tend to generalise from too small samples (Nisbett & Ross, 1980) or tend to pick the right-most of a row of identical items when asked to choose the one of superior quality (Nisbett & Wilson, 1977), as discussed in chapter 4. This knowledge can influence our psychological attributions. To put it differently, our conscious knowledge can interact with our folk psychological knowledge.

Another way of bringing out the same point is to consider how much of our knowledge of the world we bring to bear on our folk psychological attributions. In order to work out why people are doing what they are doing, we need to get at their beliefs. For example, we need to know what belief(s) can operate together with a particular desire in order to bring about the desired state of affairs. A very good guide to this are the causal connections that we have observed to hold in the past. For example, observing that stones over a certain critical size break most windows when hurled at them, I will assume that you believe the same in the absence of information to the contrary. I might draw on this knowledge when I explain what your intentions are throwing a largish stone at your ex-partner's window.

Knowledge of folk psychological theory also affects beliefs in a way quite different from that of tacit knowledge. We appear capable of directly using the principles of folk psychology in acting, forming intentions, and deciding what to make of ourselves or others. Folk psychology has often been assumed to provide us with the tools of human interaction (Fodor, 1987; Dennett, 1987). Being social animals, our thoughts and actions depend crucially on those of others. This means that, in many cases, we must take into consideration the thoughts and actions of other people in order to plan how to act to achieve our ends. Consequently, folk psychological predictions have been assumed to be the cement of human societies. Our ability to work out what other people are likely to think or do allows us to cooperate with each other.

Morton (1996) has argued that we make decisions on the basis of option-limiting procedures that are aimed at cooperation. Rather than the individual making decisions based on *predictions* of the actions or thoughts of others, the individual forms *expectations* as to the future thoughts and actions of others during or after the decision making. I cannot go into the details of Morton's idea here. Suffice it to say that I agree that there is a tendency to exaggerate the importance of folk psychological prediction for human cooperation. It is very likely that a great deal of human interaction is based on expectations, although I'm not sure whether they are a product of decisions rather than an ingredient therein. Even so, it is hard to deny that knowledge of folk psychological theory is what gives us those expectations or allows us to form them. It remains incontrovertible that at least *some* of our interactions with other people rely heavily on folk psychological prediction. I think here of manipulation, seduction, revenge, advancement, and so on, but also more benevolent actions such as planning a pleasant surprise for someone.

As an example folk psychological knowledge playing a role in decision making, think of Iago's manipulation of Othello and consequent revenge. Iago knew of the typical results of jealousy - loss of judgement and self-control, intense rage - and how it may be induced. He used this knowledge in planning his interactions with Othello. For example, he gets hold of Desdemona's handkerchief and plants it on Cassio because he believes that, on the background of the doubt he himself has already sown in Othello's mind, if Othello sees it in Cassio's hands, he will think that Desdemona gave him it as a token of her love. Here, the prediction of Othello's beliefs forms the basis of Iago's action. Old-fashioned detective stories, such as the majority of Agatha Christie's novels, have the criminal foiling the police by carrying out a number of deceptive manoeuvres to avoid detection. Such deception also relies on being able to foresee what other people will think and do under certain circumstances. I think cases such as these are best seen as involving subjects drawing directly on their folk psychological theory. Here it is not a matter of thinking if I do this, what will she do. The issue is rather that I want her to do this, and I need to know how to make her do it. It will be impracticable to go through all the different actions that one imagines one might perform under the relevant circumstances to see what one would predict that she would do. Some kind of guiding light is needed here: folk psychological theory. For example, I want to destroy what my enemy values the most - the love between him and his wife. Jealousy can destroy love, so I'll make him jealous. But in order for folk psychological knowledge to serve this role, it must be inferentially integrated.

Folk psychological theory can also be used instrumentally. It is often applied to inanimate objects, for example - objects that the attributer does not suppose to actually possess the attributed states.[57] Thus, a deciduous tree that fails to shed its leaves in autumn is easily and intelligibly described as thinking that it is still summer. Nevertheless, relatively few of those willing to attribute such a state to

[57] This fact plays a large role in Dennett's Intentional Stance Theory (1987).

a tree, believe that trees think. Rather, in the absence of arboreal knowledge, 'thinking' is used as a shorthand for whatever mechanism a tree has of gauging the season. Examples abound. I have heard a respectable astrophysicist on national television attribute intentional states to objects in space: the gasses in a quasar *try* to orbit a black hole. Tolstoy famously lamented the extension of common sense psychological explanations to governments and countries in *War and Peace*. In all of these cases we inventively apply the principles of folk psychological theory to quite disparate phenomena. This stands in sharp contrast to how we are able to use our knowledge of both transformational grammar and visual parameters. It also indicates an intimate connection to belief since when one uses a body of knowledge instrumentally, one is aware of the sense in which one applies it, in the case at hand, and how that differs from standard applications of it. Deciding to use a body of knowledge in a particular context also seems to rely on the inferential integration of that information. Therefore, there is strong evidence that folk psychological knowledge is not inferentially encapsulated.

Lastly, we must examine whether our folk psychological knowledge is non-conceptual. Firstly, there is little doubt that we possess the concepts that are involved in such knowledge. When we attribute folk psychological states, such as beliefs, desires, intentions, and actions, we do so using the very terms that are involved in the knowledge that theory theorists attribute us.[58] This, seems to be no coincidence, but due to the knowledge in question being structured by concepts. People's folk psychological knowledge seems to meet the Generality Constraint. 'Belief', 'desire', 'hope', 'fear', 'the cat is on the mat', 'the music will stop', and so on, are all thought radicals that can combine in the fashion demanded by the Generality Constraint. I can apply a desire to a tree, a moped, a worm, and so on. I can desire, believe, hope, and fear that the cat is on the mat. What we need to rule

[58] This argument is culture relative. If we assume that there is at least a core of folk psychology that is not culture specific, then we face the problem of cultures where the psychological vocabulary is significantly different from ours. I don't think this is a serious problem although it provides practical difficulties. What needs to be shown in these cases, is that subjects acknowledge the kind of psychological differences that are reflected in the range of psychological states posited by folk psychological theory. They need not have a single word for each such state. If that can be done, then we can assume that they possess the relevant concepts. Thus, they can be assumed to have conceptual folk psychological knowledge. However, for simplicity of exposition, I have chosen a more culture relative example.

out, is the possibility that I have some explicit knowledge that is causally efficacious in these combinatorial capacities that is separate from that which is causally efficacious in my normal folk psychological attributions. In other words, we need to rule out the possibility that the folk are related to their folk psychological knowledge in the way that linguists are related to their grammatical knowledge. There seem to be important discrepancies between the folk psychologist and the linguist case. Firstly, linguists are attributed knowledge of grammar prior to them behaving in ways that make it appear that their grammatical knowledge meets the Generality Constraint. In the case of folk psychological knowledge, people are attributed such knowledge at roughly time when they behave such that their knowledge seems to meet the Generality Constraint. They can conceive of a belief having as content all the propositions they have an idea of, and they can conceive of any one proposition forming the content of the variety of folk psychological states. Secondly, in the case of the linguist's knowledge meeting the Generality Constraint, we can identify a learning period in which it is plausible to suppose that she acquired the concepts and the explicit knowledge. A comparable situation cannot be found with respect of folk psychological knowledge. Even though philosophers and psychologists are trained in the area, their knowledge seems to have passed the Generality Constraint all along. This also rules out the possibility of the folk psychologist case being like the psycho-analytic one, in which psychological states come to surface in consciousness. We have no reason to think that subjects acquire folk psychological concepts separately from acquiring folk psychological knowledge. Indeed, looking at development, children appear to acquire the concepts of folk psychological theory along with the theory itself (Wellman, 1990). Knowledge of folk psychological theory goes hand in hand with folk psychological concept possession. This is what we would expect if the knowledge was conceptual.

Let me summarise what we have found so far. There is good evidence to support the claim that knowledge of folk psychological theory is consciously accessible, inferentially integrated, and conceptual. All of these characteristics are supposed to be defining characteristics of beliefs as opposed to subdoxastic states. Since beliefs mark the states of ordinary knowledge and subdoxastic states mark the states of tacit knowledge, we must conclude that folk psychological knowledge is not tacit.

6. Last Objections

Let us consider some final objections to the above conclusion. First, it might be objected that I have not ruled out the possibility that folk psychological knowledge is tacit, since subjects might have *learnt* the principles of folk psychological theory through discussion and reflection, after they have tacit mastery of them. But this case should be just like the linguist case, and we have already seen that there are important differences.

Another objection rises out of the first. Let us grant that the folk psychological generalisations considered in chapter 1, are consciously accessible, inferentially integrated, and conceptual. How do we know that *all* of folk psychological knowledge is like this? Couldn't there be tacit parts of folk psychological knowledge (cf. Scholl & Leslie, 1999)? We cannot definitively reject this possibility before we have a more elaborate formulation of the suggested part of folk psychological theory that is supposed to be tacitly known. However, as far as we know, there are no tacit parts to folk psychology, and we have been given no reason to think that there are any such.

Hang on, someone might say, the above only goes through because you have misportrayed transformational grammar. There are many things about it that we know. We have the concept of a verb, a noun, an adverb; we understand that a sentence must at least contain a subject and a verb, and so on. In short, certainly *some* of transformational grammar counts as ordinary knowledge, just like some of folk psychological knowledge does. This, however, is insufficient to show that the body of knowledge as a whole is not tacit. All I have shown is that there are parts of folk psychological theory that are explicitly known. But so are parts of transformational grammar. Therefore, I have failed to show that folk psychological knowledge is not tacit. There are several things to say in this context. A first answer would be that it is not unlikely that the explicit knowledge of grammar that we have, is what we learnt in school. Although there is a big difference between the grammar that we learn there and transformational grammar, there are certainly some similarities. These may account for the explicit knowledge that we have of grammar. Here again, we have a case of explicit learning through teaching. There is another, stronger response. This is that it is unnecessary for linguists to work with such categories as *verb*, *noun*, and so on. Categories such as these need play no role in a transformational grammar; it can do without them. It may even be that other categories are more appropriate. Now, compare this with folk psychological knowledge. How would a folk psychological theorist work without categories such as *belief*, *desire*, *hope*, and so on? They form part of what his theory must explain, because it is a theory of the practice of attributing such states. How would knowledge of a theory containing no psychological terms culminate in the attribution of them? It seems that psychological categories are much more intimately connected with an internal account of folk psychology than linguistic categories, such as *verb* and *noun*, are connected to linguistic theory. What linguistic theory must explain is people's ability to utter and comprehend grammatical sentences, and terms like 'verb' and 'noun' only play a role insofar as they occur in these sentences.

Taking a view such as this commits one to a quite strong view that it is unlikely that folk psychology can be accounted for in terms of knowledge of a theory that does not contain terms such as 'belief', 'desire', and so. Therefore, one might prefer simply to say that the difference between knowledge of transformational grammar and folk psychological theory is that what is explicitly known in the former case are small and peripheral parts of the theory, whereas what is explicitly known in the latter case form part of the core of folk psychological theory. Either way, there is a fundamental difference between folk psychological knowledge and grammatical knowledge. Only the latter is tacit.

But perhaps the above accounts of tacit knowledge are all wrong. In the future, we may come across another account of tacit knowledge that will classify folk psychological knowledge as tacit. This objection has two parts. First, there is the possibility that none of the above accounts are correct. Secondly, there is the possibility that another account of tacit knowledge will have folk psychological theory being tacitly known. Establishing the first does not establish the second. That requires additional evidence and argument. Let us begin with the first part.

Psycho-analytic states provide a *prima facie* problem for the above accounts of cognitive states. For Chomsky, there is no way of distinguishing subdoxastic states from any other consciously inaccessible states. However, psycho-analytic states seem very different from subdoxastic states. They are not posited to explain any particular ability, and they appear in a wide variety of contexts (see below). For Stich, the problem is the following. Intuitively, it may seem that his account distinguishes between psycho-analytic states and subdoxastic states because whereas they are both consciously inaccessible, only subdoxastic states are inferentially encapsulated. Psycho-analytic states manifest themselves in all aspects of everyday life: in mistakes, dreams, (certain kinds of) forgetfulness, slips of the tongue, and so on (Freud, 1900/1953). There is no one area, or smaller group of areas, in which psycho-analytic states exclusively manifest themselves. It is a mistake to think that such states are manifested only in the behaviour of the neurotic or psychotic. Hence, we cannot here find a neat parallel to the circumscribed areas in which grammatical knowledge manifests itself (linguistic utterances and comprehension). However, this cannot be Stich's view because we cannot talk about such states, and when we assent to statements about how we feel and think unconsciously, we do so not in the immediate way in which we assent to statements expressing the contents of beliefs that we have. Therefore, there is a real danger of psycho-analytic states falling on the side of subdoxastic states on Stich's classification. This, however, is infelicitous as we have seen that unconscious psychological states seem significantly different from subdoxastic ones. It might be thought that Stich has made allowances for this by pointing out that he is only speaking of normal subjects (1978, p. 505). However, as we saw above, unconscious psychological states manifest themselves in all trades of life. Everybody is subject to dreams, for example. Furthermore, everybody has unconscious psychological states even when they are not repressing (Freud, 1915/57, pp. 192-95). Therefore, it is not clear that Stich's theory is sufficient for providing a good distinction between beliefs and subdoxastic states. It may class some states as subdoxastic that shouldn't be so classified.

One might, of course, object that it is far from clear that unconscious psychological states, as presented by psycho-analytic theory, are respectable entities at all. A number of philosophers and psychologists certainly appear to think that psycho-analysis is unscientific claptrap. Alternatively, one might claim that there really is no important difference between psycho-analytic states and subdoxastic ones. That, however, would need some explanation. But if one is sympathetic to psycho-analysis broadly conceived, the above should make one hesitant to embrace Chomsky's and/or Stich's accounts without modifications.

The situation is not great for Davies either. He wants his account to classify psycho-analytic states as conceptual states. The problem here is that psycho-analysts sometimes attribute to subjects psycho-analytic states even when the knowledge is unlikely to be conceptual. Neonates can already be attributed ideas to the effect that the breast is good or bad (in Kleinian theory). Psycho-analysis is concerned with how best to explain a subjects' behaviour. If that is done by reference to unconscious states, it is no doubt useful to posit representational states that can combine in certain ways. However, whether the combinatorial capacities connected with such structures are like those connected with concepts or not, is irrelevant. Unconscious psychological states are interestingly different from ordinary psychological states. In other words, psycho-analytic states need not be conceptual states. The response to be made on Davies' behalf here is similar to the responses possible on behalf of Chomsky and Stich: Davies is misguided in paying any credence to psycho-analysis. However, the point is still worth noticing. We may put it conditionally. If there are unconscious psychological states much like psycho-analysis says there are, then none of the above accounts appear satisfactory as they stand. Some addition or reworking would be necessary.

As I said above, from the mere fact that none of the present accounts of tacit knowledge are entirely satisfactory, we cannot conclude that it may turn out that folk psychological knowledge is tacit. For what must be shown is that tacit knowledge is neither consciously inaccessible, inferentially encapsulated, nor non-conceptual. That, I believe will be hard to show, but let us have a look at some of the recent research in experimental psychology on implicit knowledge. Here, 'implicit knowledge' is used synonomously with 'tacit knowledge'.[59]

7. Kinds of Knowledge

Research into implicit knowledge is normally carried out in the context of testing what is implicitly learnt. It is an underlying assumption that implicit knowledge is implicitly learnt. The prototypical example of implicit learning and knowledge concerns artificial grammar, *not* transformational grammar. An artificial grammar is a set of relatively simple rules that guides how a finite list of elements might be combined, most commonly letters. Some letters must start a grammatical string, others end it, only some letters can follow upon other letters, and so on. In the learning phase of the experiment, subjects are exposed to so-called grammatical strings - strings of letters that are ordered according to particular rules. They are asked to memorise as many of these strings as possible. Exposure is somewhere around 5-10 seconds. In the test phase, subjects are presented with new strings, only some of which are grammatical, and are asked to classify them either as grammatical or ungrammatical. It is found that subjects perform significantly above chance. A number of researchers take this to show that the subjects have implicitly learnt the relevant rules (Manza & Reber, 1997).

Implicit knowledge is primarily characterised by it being difficult to access. Subjects possessing implicit knowledge: i. do not tend to elicit such knowledge in free recall, ii. have problems eliciting this knowledge in forced-choice tests, iii. show low confidence in their judgements, and iv. are not good at transferring their knowledge across domains (Berry & Dienes, 1993). What subjects elicit spontaneously is either not appropriate to explain their ability, or if it is, it is insufficient to do so. In general, subjects perform better at forced-choice tests than at spontaneous recall. More recent research has shown that there is some transfer across domains of implicit knowledge. G. Altman, Z. Dienes, and A. Goode (1995) found that

[59]Cf. the following passage from Mark F. St. John and David R. Shanks (1997, p. 162): "Together with its synonyms 'tacit' and 'covert', the term 'implicit' has become common coinage in psychology over the last decade. The study of implicit processes is now the focus of major research efforts in psychology, but there has been some controversy about how best to define this form of knowledge."

subjects managed to transfer knowledge of an artificial grammar across modalities: from letters to music, and from graphic symbols to nonsense syllables. Dienes & Altman (1997) found a significant transfer of knowledge of artificial grammar from words to colours. Nevertheless, although subjects perform significantly above chance on the transfer domain task, their performance is significantly worse in the new domain compared to the domain in which they were originally trained. So, although tacit knowledge may be less inferentially encapsulated than the above accounts might lead us to expect, it is still nothing like ordinary knowledge.

The only approach to tacit knowledge found in the psychological literature that differs significantly from Chomsky's, Stich's, and Davies' accounts, stems from an increasing scepticism about the abstractness of implicit knowledge. A number of researchers deny that what subjects learn are the relatively abstract rules according to which the grammars were constructed (Dienes & Altman, 1997; St. John & Shanks, 1997). Instead, they suggest that what the subjects learn are the particular configurations of the domain in which it is acquired, for example simple correlations (Perruchet & Gallego, 1997). This has lead Axel Cleeremans (1997) to deny that implicit knowledge is representational in any traditional way. It is not composed of discrete symbolic entities, but consists in "patterns of activation that are distributed over many processing elements." (p. 226). Nothing is represented separately from the processing elements. In short, this amounts to a denial that tacit knowledge is representational in the ordinary sense of that term. This, then, might be regarded as an alternative account of tacit knowledge; beliefs are representational, subdoxastic states are not. This certainly conflicts with the above accounts. However, there are reasons not to get too excited about Cleerenmans' conclusion even if we grant him, which we need not, that artificial grammatical knowledge is not representational. Artificial grammars are small and simple, containing nothing close to the complexity of transformational grammar. It may be that we use different methods of learning and storing for simple bodies of information compared to more complex ones. For example, we may need representations for the latter.

The views about tacit knowledge in the psychological literature range from subtle variations on themes suggested by, for example

Stich, to large-scale denials of the representationality of tacit knowledge. Nevertheless, there seems to be no suggestion in the offing that will allow consciously accessible, inferentially integrated, and conceptual knowledge to count as tacit. So, unless there are other kinds of knowledge that seem better candidates for subsuming folk psychological knowledge, we should conclude that it amounts to ordinary knowledge. It may be thought that folk psychological knowledge compares better to so-called expert knowledge; the knowledge that experienced chess players have of chess, doctors of diseases, physicists of physics, and so on. For, if nothing else, folk psychological knowledge seems to differ from ordinary knowledge by being more difficult to express and by subjects normally being unaware of using it whenever they do so. This is just the situation experts find themselves in. They find it difficult to verbally express their knowledge. However, they are likely to be able to recall their knowledge when given sufficient time and incentive. But there are also differences. For example, transfer of knowledge between the original and new domains is limited (Dienes & Altman, 1997). Expert knowledge appears inferentially encapsulated. Hence, folk psychological knowledge doesn't quite fit the profile of expert knowledge either. Therefore, it seems fair for us to conclude that folk psychological knowledge classifies as ordinary knowledge. It has more things in common with ordinary knowledge than with either tacit or expert knowledge.

Chapter 6

Conclusion

This is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning. (Sir Winston Churchill, *Speech, Mansion House, 10 Nov. 1942*)

In the preceding chapters, I have dealt with what I have seen as foundational issues in the Theory Theory of folk psychology. The need for philosophical foundations for this theory has been exacerbated by the Theory Theory versus Simulation Theory debate. In this debate it is apparent that there is little consensus about the nature and commitments of the Theory Theory; not even among theory theorists themselves. Consequently, parts of the debate have been somewhat misguided. The aim of the thesis has been to clear up at least some of these exegetical mistakes.

Theory Theory claims that knowledge of a folk psychological theory is causally efficacious in the production of our psychological attributions. It holds that such a knowledge is *necessary* for such attributions but not sufficient. I take it that this is a statement of Theory Theory that all who call themselves theory theorists can agree about. I have been concerned with fleshing out this claim. Three questions immediately present themselves: what does it mean to say that we have knowledge of a *theory*, what is the nature of our *knowledge* of this theory, and if knowledge of such a theory is necessary involved in attributing psychological properties to people, does this imply attributional symmetry, and that we have no distinct knowledge of our own minds, as opposed to those of others? The answer to the first is that what we know is structured in a particular way that is importantly similar to the way scientific theories are structured. The answer to the second is that our knowledge of this theory is not tacit, but like ordinary knowledge. The answer to the third is that although some attributional symmetry is required, that symmetry isn't complete. We are not required to deny that we have no distinctive knowledge of our own psychological states, indeed it seems folly to do so. This knowledge does not, however, derive from direct awareness of psychological states as we conceive of them. To put it differently, what we attribute ourselves has more to it than what we are directly presented with in experience.

A presupposition for answering the three questions is that one has a pretty solid idea of what folk psychological theory consists in. I have given a number of examples of folk psychological generalisations that I believe are uncontroversial. Many of them are mentioned in the literature in one form or other. Equipped with these, the three questions become answerable. In chapter 5, I pointed out that there are different views of what the folk psychological generalisations we have knowledge of are like, but no examples thereof. Should such examples appear, and should they turn out to satisfactorily explain some of our psychological attributions, some of what I have said in the above could be rendered obsolete. For example, the questions of whether the parallel with scientific theories can be upheld and whether the theory is tacitly known, will be reopened. However, much of substance will remain. For example, if what we know turn out to be very different in structure from scientific theories, we should resist calling our account of folk psychology 'Theory Theory'.

Throughout, I have aimed at being as undogmatic as possible, in order not to commit Theory Theory to something that it need not be committed to. However, I have closed options too. I have rejected a weak reading of 'theory' with the consequence that one theory that calls itself a Theory Theory is excluded from being so. This is unfortunate, but unavoidable. In order to provide a solid basis for Theory Theory, we must make its claim relatively precise. On the other hand, we may also end up appropriating theories that don't call themselves Theory Theories, but share some of the same basic presuppositions. This was the fate of Davies & Stone's version of simulationism. Firm answers to the three questions mentioned seems to me necessary to have a foundation on which to build a properly worked out Theory Theory. If we do not provide such answers, we are stranded with an impossibly vague position. Instead, once we accept the picture of Theory Theory presented here, we have something much more specific to deal with. It will allow us to consider whether Theory Theory is a reasonable position as it stands and, if not, what other position might best capture our ideas.

Wanting some firm ground on which to construct Theory Theory, I nevertheless attempted to leave a number of issues open for debate among theory theorists. These are the following. A theory theorist may be a metaphysical or a semantic common sense functionalist. What I have pointed out is simply that Theory Theory is not synonymous with functionalism. All functionalism requires is some systematisation of folk psychological generalisations, but can be content with an external account of folk psychology. Theory Theory, however, is specifically an internal account. If one desires to say something about the meaning of psychological state terms and about the nature of psychological states, one can embrace functionalism as well as Theory Theory. But this is far from necessary. It is worth keeping in mind that functionalism is not an unproblematic position (Block, 1980). Fixing the reference of theoretical terms exclusively in terms of the role that they play in a theory, has the counterintuitive consequence that there can be no trans-theoretical terms. Take the term 'electron'. It is a term that has been used in a number of different theories. Niels Bohr used it, for example, but the prevalent theory in which the term figures now, is different from Bohr's. If we adopt a functionalist theory of the meaning of theoretical terms, we commit ourselves to maintaining that what Bohr meant by 'electron' was different from what *we* mean by 'electron'. We are not talking about the same thing. Consequently, it will be false to say that Bohr was wrong about electrons. He *couldn't* be wrong because he was concerned with something different from electrons, that we now know don't exist. Functionalists try to obviate this difficulty by saying that if theories are a bit false, this does not influence the meaning of the terms. The problem remains of determining just how false one can allow a theory to be before there is a change in terms. And it seems to me that common sense has it that we can talk of a theory being quite wrong about electrons as opposed to a theory having a different notion of electrons (say, Bohr's). This, then, is one problem that functionalists face.

Another option that is left open for the theory theorists is just what account of self-attribution and self-knowledge is deemed to be most satisfactory. The options are not unlimited. Full-scale symmetric or asymmetric positions cannot be chosen, and there are the empirical data to be accounted for also. Nevertheless, Theory Theory seems compatible with a number of current accounts, either as they stand or slightly revised versions of them. By keeping options such as these open, Theory Theory will remain a precise and distinctive internal account of folk psychology, but not thereby firmly committed on a number of other issue. These issues *are* relevant to Theory Theory, but not relevant in the sense theory theorists are required to take a particular stand with respect of them in order to count as theory theorists. We must keep open a window for disagreement. But we must make sure that such disagreement does not concern the foundations of the theory.

I have only been able to deal with just a few issues surrounding Theory Theory. I take it, however, that these were the most pressing to get resolved. I leave a host of questions unanswered. Below I will hint at other foundational issues that are important for theory theorists to deal with. There are three issues, I would like to pick up on. They concern the acquisition of folk psychological theory, *ceteris paribus* clauses, and the compilation of a properly explanatory folk psychological theory.

1. Acquisition and Development of Folk Psychological Theory

Much work on the Theory Theory has been done by child psychologists. In chapter 1, we were introduced to two accounts of the acquisition of folk psychological theory. There is Gopnik, Meltzoff, and Wellman's view that children are little scientists, or, alternatively, that scientists are big children, with an innately given ability to theorise. Children acquire folk psychological theory by forming a theory about the contingencies that they observe in the world. We were also introduced to Baron-Cohen's work, inspired by Leslie, on the precursors of folk psychological theory. The ID, EDD, and SAM come on-line at various stages of development, and provide input to the ToMM. However, this says nothing about how one develops folk psychological theory once these modules are on-line.

Embracing a less narrow reading of 'theory' than that suggested by Gopnik, Meltzoff, and Wellman might lead one to embrace a less scientistic view of child development. Nevertheless, if it seems that what children learn is a theory on the very strict notion defended by Gopnik, Meltzoff, and Wellman, the same evidence can be used to support the more liberal version that I have defended. If we accept a more liberal idea of theoreticity, greater discrepancies between what scientists do in their labs and what children do in their homes are allowed. Those discrepancies merit more attention. Indeed, I think it is crucial not to overemphasize the ways in which children and scientists are alike. There are obvious differences between acquiring folk psychological theory and being a scientist. Theory Theory must allow for these. I am not claiming that psychologists deny that there are differences between scientists and children, but rather than people like Gopnik have a tendency to marginalise them.

The means by which people acquire folk psychological theory are not unlikely to change over the years. Young children seem to behave in ways that are comparable to experimentation. They repeat certain actions over and over, in order, it seems, to observe whether the same effect follows: "Will mother shout when I turn over my plate (again)?" Alt-explan? As children grow bigger, they become increasingly unlikely behave in to this way. Behaviours that are tolerated in young children are not tolerated in older children. Being told off and meeting hostility and anger are hardly conducive of freewheeling experimentation. As children become adolescents, the situation becomes more difficult. Although Gopnik seems to think that teenagers experiment,[60] I don't think that their behaviour is best seen as that. Their lack of regard for parental authority, customs, and so on, is better understood in terms of becoming more autonomous human beings, not in terms of Could be developing folk psychological knowledge. both Even teenagers have regard for their parents and do not experiment on them in order to find support or to falsify their psychological hypotheses. It is even more unlikely that people experiment with their peers. Human relationships are fragile and fraught with difficulties as it is. In general, one does one's best to get on. People agonise over having said or done the right thing because they know just how severe punishments can follow from doing or saying the wrong thing. Here is no joyful experimenting when what could be at stake is loss of job, position, dear ones, and so on.

What I was trying to emphasize in chapter 1, was that it is possible to regard any one person's folk psychological knowledge as continually developing. The development in the first five or ten years is no doubt much more dramatic than that which follows it. Nevertheless, we continue to learn new things about people. Theory Theory's account of acquisition of folk psychological theory should account for this. As I have been at pains to point out, modelling such development on scientific experimentation is not always very plausible. For example, it is quite clear that people frequently generalise from their own case, in the absence of further supporting evidence. If you

[60] Private conversation.

have a tendency to dislike being corrected, the temptation is assume that everybody dislikes being corrected. However, some people like being corrected because they believe that it will improve their performance. Cases like these are clearly very different from cases of scientists developing theories. Therefore, it would be useful for Theory Theory to be more specific about the different ways in which we acquire folk psychological theory, and how these ways may change over the years, as we become adults. I'm not sure this is really a foundational issue, but it is certainly an issue that needs to be addressed before Theory Theory is likely to recruit new supporters.

2. Ceteris Paribus

A serious challenge for the theory theorist is to explain how we manage to represent *ceteris paribus* clauses. One might grant that our folk psychological theory is explicitly known, but if one adds the *ceteris paribus* clauses to each and every generalisation, we seem to end up with a ridiculous amount of information. However, in many situations, we effortlessly attribute psychological states and we must, therefore, master such knowledge easily and quickly. How is this possible if what we need to do is to consult a gargantuan body of knowledge?

One option to take is to deny that adding *ceteris paribus* clauses to folk psychological generalisations makes the amount of information unmanageable. This is not as crazy as it sounds. We know very little about how we manage to use the relevant parts of any of our larger bodies of knowledge in particular situations. Most of them are somewhat complex, including a substantial number of *ceteris paribus* clauses. We need to know how we manage to use the relevant information as effortlessly and quickly as we do. It is unlikely to turn out to be the case that the normal situation is extremely simple, but that the folk psychological one is impossibly complicated. Furthermore, although there are a number of conditions where *ceteris* are not *paribus*, it is somewhat pessimistic to imagine that a long list is attached to each folk psychological generalisation.

What seems to be another option is to liken folk psychological knowledge to expert knowledge. This would account for our ease at applying our knowledge in certain cases. However, we saw in chapter 5 that expert knowledge is significantly different from folk psychological knowledge. It is quite unclear that there is any class of knowledge other than ordinary knowledge that the latter fall under. But perhaps much of what we regard as ordinary knowledge is automated in some manner. After all, subjects may have *some* awareness of their automated knowledge, but need not know exactly how it is causally efficacious in the production of any given behaviour. This might account for our use of folk psychological knowledge. Nevertheless, rather than being an alternative to the first option, it *is* the first option. The problem with this particular version of the first option, is that it is not clear that subjects cannot report on the knowledge that is causally efficacious in their psychological attributions.

It is, I believe, rather important for Theory Theory to be clearer on the issue concerning the application of folk psychological theory. This does tie closely in with the application of larger bodies of knowledge in general. Possibly no resolution will be found until we know more about knowledge and the application of it. However, there is still room for some development of the Theory Theory stand on this issue.

4. Compiling Folk Psychological Generalisations & Psychologic

What I presented in chapter 1 was but a handful of generalisations. However, consider the following possibility. When we consider our folk psychological practice, there is a handful of psychological generalisations that naturally present themselves. This gives rise to the idea that knowledge of a body of generalisations such as these, is what is causally efficacious in our psychological attributions. However, it might turn out that when we turn to look for more such generalisations, in order that we can explain a more substantial part of our behaviour, we realise that we were mistaken. We cannot come up with any more such generalisations, at least not enough to support Theory Theory. I believe this is an important objection for theory theorists to counter. Not, of course, by providing a complete folk psychological theory, but by providing more generalisations to substantiate the position.

Now, it so happens that a psychologist has done just that. Jan Smedslund has long been engaged in carrying out the project of listing the generalisations that form the core of folk psychological theory (Smedslund, 1990, 1997). He sees himself as: (1997, p. ix)

> explicating the implicit conceptual system of psychology embedded in ordinary language, or in other words, the basic assumptions and distinctions underlying our ways of thinking and talking about psychological phenomena.

ext

int?

He calls such an explication 'psychologic'. The idea behind psychologic research is to unveil the "invariant structure embedded in the way we talk and think about persons, and deal with them." (1990, p. ix). He takes this as being a foundational issue in psychology. Psychology uses psychological terms unreflectively. In order that psychology may progress as a truly scientific discipline, the meaning of terms deployed by this discipline must be laid bare. In this way, scientific psychology can be based on an explicated folk psychological theory.

Smedslund's view of folk psychological theory is quite distinctive. Firstly, he believes that folk psychological generalisations are normative, not descriptive, statements: (1990, p. 60)

> psychologic owes its predictive success to its being an explication of rules which people regard as correct and live according to. These rules are man-made and maintained by people, and hence are very different from natural laws.[61]

?

However, as I argued when I converted Rational Simulationism into Rational Theory Theory, working with normative generalisations will not, by itself, disqualify an internal account of folk psychology from being a Theory Theory. Secondly, he does not believe that folk psychological theory is an empirical theory. He believes that is an *a priori* theory, because if you substitute the occurrence of a psychological term in a psychological law with the definition of it, the

[61] There is a tension between this way of conceiving of folk psychological theory, and Smedslund's conviction that his psychologic is nothing but a system of Lewisian platitudes (1997, p. xii). Lewis clearly does not hold that folk psychological platitudes are normative, but that they are descriptive. Furthermore, Lewis does not regard the platitudes, themselves, as being causally efficacious in the production of our folk psychological attributions. Smedslund does.

law becomes tautological.[62] This, of course, is an artefact of the functionalist analysis of the meaning of theoretical terms. One could say that the theory is nothing but an explication of some of the terms involved in it - in this case, folk psychological theory explicates the meaning of folk psychological terms. Nevertheless, it remains an empirical fact whether or not the theory applies to anything. The meaning of theoretical terms is analytic, but it is an empirical question whether there is anything to which they apply. All this, however, seems perfectly in accordance with Smedslund's idea. He, himself, however, has failed to see that once you apply a Ramsey-Carnap-Lewis theory of the meaning of theoretical terms to a theory, that theory will turn out to be just as *a priori* as folk psychological theory. Smedslund thinks this is an artefact of certain theories only, geometry being the prototype. Thirdly, he seems to think that knowledge of folk psychological theory is implicit which, if the general use in psychology of this term is anything to go by, means that he thinks that it is tacit. Looking ahead at (1)-(13), however, I fail to be convinced. However, the point here is not to examine Smedslund's project, but to look at some of the results of this project: the list of folk psychological generalisations.

In *The Structure of Psychological Common Sense*, Smedslund presents his psychologic as a sort of geometrical treatise, divided into definitions, axioms, theorems, corollaries, and explanatory notes. This appears to be connected to his idea that folk psychological theory is like Euclidean geometry; both are composed of logically necessary propositions and both are useful tools for prediction (1990, p. 45). There are 56 axioms. Here is a handful: (1997, p. 104-7, the numbering is mine)

> (1) A person is held responsible for his or her acts by everyone involved.

[62] The example he gives concerns surprise. The definition of surprise is: " '*Person* P *in situation* S *at time* t *is surprised*' = df 'P *in* S *at* t *is in a state of having experienced something that* P *had expected or had taken for granted would not occur.*' " (1990, p. 48). The relevant law is described as follows: " '*If* P *in* S *at* t *experiences an event which* P *has expected, or taken for granted, not to occur, then* P *in* S *at* t *will become surprised.*' " (p. 54). Inserting the definition of 'surprise' into the law produces the following result: " '*If* P *in* S *at* t *experiences an event which* P *has expected or taken for granted would not occur, then* P *in* S *at* t *will be in a state of having experienced something which* P *had expected or had taken for granted would not occur.*' " (p. 55). This is clearly a tautology.

(2) *P* wants to do what *P* believes is right and to reject what *P* believes is wrong. ?

(3) A conscious person is continuously acting. ?

(4) *P* tries to maximize expected utility. ?

(5) *P* wants to feel good and wants to avoid feeling bad. ?

(6) *P's* want *A* is stronger than *P's* want *B*, if, and only if, when *A* and *B* are in conflict, and no other factors intervene, *P* tries to act according to *A* and not according to *B*.

(7) *P* wants to believe what is the case. ?

(8) If everyone takes a psychological proposition *X* to be self-evident, then everyone believes that everyone else takes *X* to be self-evident, everyone believes that everyone else believes that everyone else takes *X* to be self-evident, everyone believes that everyone else believes that everyone else believes that everyone else takes *X* to be self-evident, and so on. ?

(9) The strength of a feeling is equal to the product of the strength of the want and the strength of the belief, whose relationship constitutes the feeling. ?

(OS theory of emotio

(10) Every person wants to care for someone. ?

As I mentioned, these are only axioms. For each axiom there are varying number of theorems, corollaries, and notes. For example: (p. 63-4)

> (11) Every person wants to be cared for by someone, [and]
>
> (12) *P* wants his or her liking a person to be reciprocated.

? are theorems of (10). An example of a note to (10) is: (p. 63)

> (13) The preceding axiom does not state that persons always care for someone or act caringly. It only asserts that a want to care for someone always exists. Whether it is manifested in action depends on its strength relative to other wants. ?

Now, I'm not too sure about applying the structure of a geometrical treatise to folk psychological theory. However, we need not get embroiled in the logic of psychologic, what is important is simply the generalisations listed. They form a good starting point for a theory theorist. Some of the generalisations are more satisfying than others, no doubt.[63] The point is not to advocate Smedslund's psychologic, but to point out that there has been a substantial attempt at fleshing out folk psychological theory. And whatever you might want to say about (1)-(13), they are not obviously hopeless candidates for folk psychological generalisations.

!
Ya reelen?

5. The End

Let me, somewhat perversely, end this conclusion with a conclusion. I have dealt with just a few foundational issues in Theory Theory. There are many other such issues that need addressing, three of which I have briefly mentioned above. Providing satisfactory solutions to the problems raised is necessary for Theory Theory to constitute an attractive and plausible internal account of folk psychology. I had a brief look at the possible directions in which such research might take us. I think you will agree that the prospects for Theory Theory are not bad. It may ultimately not turn out to be the right internal account of folk psychology; most people nowadays seem happy about a Theory Theory-Simulation Theory mix. However, I hope to have shown in the above that, as it stands at the moment, Theory Theory is a live option.

[63] I, for one, have misgivings about the formulation, but not the idea of (4). If (4) is accepted as a folk psychological generalisation, it seems a candidate for a generalisation that is tacitly known.

Bibliography

Altman, G., Dienes, Z. & Goode, A.: (1995) "Modality Independence of Implicitly Learned Grammatical Knowledge", *Journal of Experimental Psychology: Learning, Memory, and Cognition 21*, pp. 899-912

Anscombe, G. E. M.: (1957) *Intention*, Blackwell, Oxford

Aquinas, T.: (1989) *Summa Theologiae: A Concise Translation*, (ed.) McDermott, T., Methuen, London

Aristotle (1976) *The Nicomachean Ethics*, Penguin Books Ltd, Harmondsworth

Armstrong, D.: (1993) "Causes are perceived and introspected", *Behavioral and Brain Sciences 16*, p. 29

(1994) "Introspection", in Cassam, Q.: (ed.) *Self-Knowledge*, Oxford University Press, Oxford, pp. 109-117

Astington, J. (1994) *The Child's Discovery of the Mind*, Fontana, London

Astington, J. & Gopnik, A.: (1988) "Knowing you've changed your mind: Children's understanding of representational change", in: Astington, J., Harris, P. & Olson, D.: (eds) *Developing Theories of Mind*, Cambridge University Press, Cambridge, pp. 193-206

Atran, S. (1994) "Core domains versus scientific theories: Evidence from systematics and Itza-Maya folkbiology", in: Hirschfeld, L.A. & Gelman, S.A.: (1994), pp. 316-40

Baars, B. J.: (1988) *A Cognitive Theory of Consciousness*, Cambridge University Press, Cambridge

Baron-Cohen, S.: (1995) *Mindblindness*, MIT Press, Cambridge, MA

Baron-Cohen, S., Jolliffe, T., Mortimer, C., & Robertson, M.: (1997) "Another Advanced Test of Theory of Mind: Evidence from Very High Functioning Adults with Autism or Asperger Syndrome", *Journal of Child Psychology and Psychiatry 38*, pp. 813-22

Baron-Cohen, S., O'Riordan, M., Jones, R. Stone, V. & Plaistead, K.: (1999) "A new test of social sensitivity: Detection of faux pas in normal children and children with Asperger syndrome" *Journal ofAutism and Developmental Disorders 29*, pp. 407-18

Berry, D. C. (1997) (ed.) *How Implicit is Implicit Learning?*, Oxford University Press, Oxford

Berry, D. C. & Dienes, Z.: (1993) *Implicit Learning*, Lawrence Erlbaum Associates, Hove

Block, N.: (1980) "Troubles with Functionalism" in his: (ed.) *Readings in the Philosophy of Psychology*, Methuen, London, pp. 268-305

Braddon-Mitchell, D. & Jackson, F.: (1996) *Philosophy of Mind and Cognition*, Blackwell, Oxford

Burge, T.: (1996) "Our Entitlement to Self-Knowledge", *Proceedings of the Aristotelian Society XCVI*, pp. 91-116

Campbell, R. L. & Bickhard, M. H.: (1993) "Knowing levels and the child's understanding of mind", *Behavioral and Brain Sciences 16*, pp. 33-4

Carey, S.: (1985) *Conceptual Change in Childhood*, MIT Press, Cambridge, MA

Carnap, R.: (1956) *Meaning and Necessity*, Chicago University Press, Chicago, IL

Cartwright, N.: (1983) *How the Laws of Physics Lie*, Clarendon Press, Oxford

(1989) *Nature's Capacities and their Measurement*, Oxford Univeristy Press, Oxford

Cherniak, C. (1986) *Minimal Rationality*, MIT Press, Cambridge, MA

Chomsky, N.: (1975) *Reflections on Language*, Pantheon, New York

(1986) *Knowledge of Language*, Praeger Publishers, New York

Churchland, P.: (1970) "The Logical Character of Action-Explanations", *The Philosophical Review 79*, pp. 214-36

(1979) *Scientific Realism and the Plasticity of Mind*, Cambridge University Press, Cambridge

(1981) "Eliminative Materialism and the Propositional Attitudes", *Journal of Philosophy* 78, pp. 67-90, reprinted in: Goldman, A. (ed.): *Readings in Philosophy and Cognitive Science*, MIT Press, Cambridge, MA

(1988) "Folk Psychology and the Explanation of Human Behaviour", *Proceedings of the Aristotelian Society, suppl. vol.*, pp. 209-21

Cleeremans, A.: "Principles for implicit learning", in Berry, D.C. (1997), pp. 195-234

Crane, T.: (1995) *The Mechanical Mind*, Penguin Books, London

(1998) "Intentionality as the Mark of the Mental", in: O'Hear, A.: (ed.) *Contemporary Issues in the Philosophy of Mind*, Cambridge University Press, Cambridge, pp. pp. 229-52

Darwin, C: (1872/1994) *The Origin of the Species*, Senate, London

Davidson, D.: (1963) "Actions, Reasons, and Causes", *Journal of Philosophy 60*, pp. 685-700

(1970) "Mental Events", in: Foster, L. & Swanson, J. W.: (eds.) *Experience and Theory*, University of Massachusetts Press and Duckworth, MA., pp. 79-101, reprinted in his: *Actions and Events*, Clarendon Press, Oxford, 1980

(1974) "Psychology as Philosophy", in: Brown, S. C.: (ed.) *Philosophy of Psychology*, Macmillan Press & Barnes, Noble, Inc., pp. 41-52, reprinted in his: *Actions and Events*, Clarendon Press, Oxford, 1980

(1975) "Thought and Talk", *Mind and Language: Wolfson College Lectures, 1974*, Clarendon Press, Oxford

Davies, M.: (1989) "Tacit Knowledge and Subdoxastic States", in: George, A.: (ed.)*Reflections on Chomsky*, Blackwell, Oxford, pp. 131-52

(1994) "The Mental Simulation Debate", *Proceedings of the British Academy 83*, Oxford University Press, Oxford, pp. 99-127

Davies, M. & Stone, T.: (1995a) (eds.) *Folk Psychology*, Blackwell, Oxford

(1995b) (eds.) *Mental Simulation*, Blackwell, Oxford

(1998) "Folk Psychology and Mental Simulation", in: O'Hear, A.: (ed.) *Current Issues in the Philosophy of Mind*, Cambridge University Press, Cambridge, pp. 53-82

Dennett, D. C.: (1987) *The Intentional Stance*, MIT Press, Cambridge, MA

Dienes, Z. & Altman, G.: (1997) "Transfer of implicit knowledge across domains: How implicit and how abstract?", in: Berry, D.C. (1997), pp.107-23

diSessa, A. A.: (1988) "Knowledge in Pieces", in: Forman, G & Pufall, P.: (eds.) *Constructivism in the computer age*, Lawrence Erlbaum, Ass., Hillsdale, NJ, pp. 49-70

Ducasse, C. J.: (1926) "On the Nature and Observability of the Causal Relation", *Journal of Philosophy 23*, pp. 57-68, reprinted in Sosa, E. & Tooley, M.: (eds.) *Causation*, Oxford University Press, Oxford, 1993

Evans, G.: (1982) *Varieties of Reference*, Clarendon Press, Oxford

Fodor, J. A.: (1975) *The Language of Thought*, Harvard University Press, Cambridge, MA

(1981) "The Appeal to Tacit Knowledge in Psychological Explanations", in his: *Representations*, Harvester Press Ltd., Brighton, pp. 63-78

(1987) *Psychosemantics*, MIT Press, Cambridge, MA

(1992) "A theory of the child's theory of mind", *Cognition 44*, pp. 283-96

Freud, S.: (1900/1953) *The Interpretation of Dreams*, in: *The Standard Edition of the Complete Works of Sigmund Freud, vol. IV & V*, The Hogarth Press, London

(1915/1957) *The Unconscious*, in: *The Standard Edition of the Complete Works of Sigmund Freud, vol. XIV*, The Hogarth Press, London

Friedman, M.: (1974) "Explanation and Scientific Understanding", *Journal of Philosophy 71*, pp. 5-19

Frith, U.: (1989) *Autism*, Blackwell, Oxford

Furnham, A. (1987) "The proverbial truth: contextually reconciling the truthfulness of antinomous proverbs", *Journal of Language and Social Psychology 6*, pp. 49-55

Goldman, A. I.: (1989) "Interpretation Psychologized", *Mind and Language 4*, reprinted in Davies, M. & Stone, T. (1995a)

(1993) "The Psychology of Folk Psychology", *Behavioral and Brain Sciences 16*, 15-28

(1995) "Empathy, Mind and Morals", in: Davies, M. & Stone, T. (1995b)

Gómez, J. C.: (1994) "Shared attention in ontogeny and phylogeny: SAM, TOM, Grice, and the great apes", *Cahiers de Psychologie Cognitive/Current Psychology of Cognition 13*, pp. 590-98

Gopnik, A.: (1988) "Conceptual and Semantic Development as Theory Change: The Case of Object Permanence", *Mind and Language 3*, pp. 197-216

(1993a) "How We Know Our Minds: The Illusion of First-person Knowledge of Intentionality", *Behavioral and Brain Sciences 16*, pp. 1-14

(1993b) "Theories and Illusions", *Behavioral and Brain Sciences 16*, pp. 90-100

Gopnik, A. & Astington, J. W.: (1988) "Children's Understanding of Representational Change and Its Relation to the Understanding of False Belief and the Appearance-Reality Distinction", *Child Development 59*, pp. 26-37

Gopnik, A. & Meltzoff, A. N.: (1997) *Words, Thoughts, and Theories*, MIT Press, Cambridge, MA

Gopnik, A. & Slaughter, V.: (1991) "Young children's understanding of changes in their mental states", *Child Development 62*, pp. 98-110

Gopnik, A. & Wellman, H. M.: (1992) "Why the Child's Theory of Mind Really Is a Theory", *Mind and Language 7*, pp. 145- 171, reprinted in: Davies, M. & Stone, T. (1995a)

(1994) "The Theory Theory", in: Hirschfeld, L.A. & Gelman, S.A.: (1994), pp. 257-93

Gordon, R. M.: (1992a) "The Simulation Theory: Objections and Misconceptions", *Mind and Language* 7, pp. 11-34, reprinted in: Davies, M. & Stone, T. (1995a)

(1992b) "Reply to Stich and Nichols", *Mind and Language* 7, pp. 87-97, reprinted in: Davies, M. & Stone, T. (1995a)

(1995) "Simulation Without Introspection", in: Davies, M. & Stone, T. (1995b), pp. 53-67

Happé, F. G. E.: (1994) "An Advanced Test of Theory of Mind: Understanding of Story Characters' Thoughts and Feelings by Able Autistic, Mentally Handicapped and Normal Children and Adults", *Journal of Autism and Developmental Discorders 24*, pp. 129-54

(1995) "The Role of Age and Verbal Ability in the Theory of Mind Task Performance of Subjects with Autism", *Child Development 66*, pp. 843-55

Harman, G.: (1999) " Moral Philosophy Meets Social Psychology", *Proceedings of the Aristotelian Society XCIX*, pp. 315-31

Harris, P. L.: (1994) "Thinking by children and scientists: False analogies and neglected similarities", in: Hirschfeld, L. A. & Gelman, S. A. (1994)

Heal, J.: (1986) "Replication and Functionalism", in J. Butterfield: (ed.) *Language, Mind and Logic*, Cambridge University Press, Cambridge, pp. 135-150, reprinted in Davies, M. & Stone, T. (1995a)

(1994a) "Moore's Paradox: A Wittgensteinian Approach", *Mind 103*, pp. 5-24

(1994b) "Simulation vs. Theory Theory: What is at Issue?", *Proceedings of the British Academy 83*, Oxford University Press, Oxford, 129-44

(1995) "How to Think About Thinking", in: Davies, M & Stone, T. (1995b), pp. 33-52

((1996) "Simulation and Cognitive Penetrability", *Mind and Language 11*, pp. 44-67

(1998) "Understanding Other Minds from the Inside", in O'Hear, A.: (ed.) *Current Issues in Philosophy of Mind*, Cambridge University Press, Cambridge, pp. 83-99

Hempel, C. G.: (1965) *Aspects of Scientific Explanation*, The Free Press, New York, NY

Hirschfeld, L.A. & Gelman, S.A.: (1994) (eds.) *Mapping the Mind*, Cambridge University Press, Cambridge

Horgan, T & Woodward, J.: (1990) "Folk Psychology is Here to Stay", reprinted in: Lycan, W.G.: (ed.) *Mind and Cognition*, Blackwell, Oxford, pp. 399-420

Hume, D.: (1777/1975) *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, reprinted by Oxford University Press, Oxford

Kahneman, D. & Tversky, A.: (1982) "On the Psychology of Prediction", in: Kahneman, D., Slovic, P., and Tversky, A.: (eds.) *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge, pp. 48-68

Kant, I. (1785/1993) *Grounding for the Metaphysics of Morals*, (transl.) Ellington, J. W., Hackett Publishing Company, Inc., Indianapolis, IN

Keil, F. C.: (1994) "The birth and nurturance of concepts by domains: The origins of concepts of living things" in: Hirschfeld, L.A. & Gelman, S.A.: (1994), pp. 234-54

Kripke, S.: (1980) *Naming and Necessity*, Blackwell, Oxford

Kuhn, T. S.: (1970a) *The Structure of Scientific Revolutions*, 2nd edition, University of Chicago Press, Chicago, IL

(1970b) "Reflections on my Critics", in: Lakatos, I. & Musgrave, A.: (eds.) *Criticism and the Growth of Knowledge*, Cambridge University Press, Cambridge, pp. 231-78

(1982) "Commensurability, Comparability, Communicability", in Asquith, P. & Nickles, T.: (eds.) *PSA 1982*, Philosophy Association, East Lancing, MI, pp. 669-88

Kühberger, A. Perner, J., Schulte, M. & Leingruber, R.: (1995) "Choice or No Choice: Is the Langer Effect Evidence against Simulation?", *Mind and Language 10*, pp. 423-36

Langer, E.: (1975) "The Illusion of Control", *Journal of Personality and Social Psychology 32*, pp. 311-28

Larson, R. & Segal, G.: (1995) *Knowledge of Meaning*, MIT press, Cambridge, MA

Leslie, A. M.: (1987) "Pretense and Representation: The Origins of "Theory of Mind"", *Psychological Review 94*, 412-26

(1994) "*Pretending* and *Believing:* Issues in the theory of ToMM", *Cognition 50*, pp. 211-38

Lewis, D.: (1966) "An Argument for the Identity Theory", *Journal of Philosophy LXIII*, pp. 17-25, reprinted in his: *Collected Papers Vol. I*, Oxford University Press, Oxford, 1983

(1972) "Psychophysical and Theoretical Identifications", *Australasian Journal of Philosophy 50*, pp. 249-58, reprinted in: Block, N.: (ed.) *Readings in the Philosophy of Psychology*, Methuen, London, 1980

(1994) "Reduction of Mind" in: Guttenplan, S.: (ed.) *A Companion to the Philosophy of Mind*, Blackwell, Oxford, pp. 412-31

Lipton, P.: (1991) *Inference To the Best Explanation*, Routledge, London

Loar, B.: (1993) "Functionalism can explain self-ascription" *Behavioral and Brain Sciences 16*, pp. 58-60

Manza, L. & Reber, A. S.: (1997) "Representing artificial grammars: Transfer across stimulus forms and modalities", in: Berry, D.C. (1997), pp. 73-106

Marr, D.: (1982) *Vision*, W.H. Freeman and Co., New York

Martin, M. G. F.: (MS) "Desire in Time"

Masterman, M. (1970) "The Nature of a Paradigm", in Lakatos, I. & Musgrave, A.: (eds.) *Criticism and the Growth of Knowledge*, Cambridge University Press, Cambridge, pp. 59-89

McCloskey, M.: (1983) "Naive Theories of Motion", in Gentner, D. & Stevens, A.L.: (eds.) *Mental Models*, Lawrence Erlbaum Ass., Hillsdale, NJ, pp. 299-324

McDowell, J.: (1978) "Are Moral Requirements Hypothetical Imperatives?", *Proceedings of the Aristotelian Society, suppl. vol.*, pp. 13-29

Melden, A. I.: (1960) "Willing", *Philosophical Review 69*, pp. 475-84

Morton, A. (1980) *Frames of Mind*, Oxford University Press, Oxford

(1996) "Folk Psychology is not a Predictive Device", *Mind 105*, pp. 119-37

Nagel, E.: (1961) *The Structure of Science: Problems in the Logic of Scientific Explanation*, Harcourt, Barce & World Inc., New York

Nagel, T: (1986) *The View From Nowhere*, Oxford University Press, Oxford

Newton-Smith, W. H.: (1981) *The Rationality of Science*, Routledge & Kegan Paul, Boston, MA

Nichols, S., Stich, S., Leslie, A. & Klein, D.: (1996) "Varieties of Off-Line Simulation", in: Carruthers, P. & Smith, P.K.: (eds.) *Theories of Theories of Mind*, Cambridge University Press, Cambridge, pp. 39-74

Nisbett, R. E. & Ross, L.: (1980) *Human Inference: strategies and shortcomings of social judgment*, Prentice-Hall Inc., Englewood Cliffs, NJ

Nisbett, R. E. & Wilson, T. D.: (1977) "Telling More Than We Can Know: Verbal Reports on Mental Processes", *Psychological Review 84*, pp. 231-59

Peacocke, C.: (1992) *A Study of Concepts*, MIT Press, Cambridge, MA

(1998) "Conscious Attitudes, Attention, and Self-Knowledge", in: Wright, C., Smith, B. & Macdonald, C.: (eds.) *Knowing Our Own Minds*, Clarendon Press, Oxford, pp. 63-98

Peirce, C. S.: (1933) *Collected Papers of Charles Sanders Peirce, Vol. V*, C. Hartshorne & P. Weiss (eds.), Harvard University Press, Cambridge, MA

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

Perner, J., Ruffman, T. & Leekam, S. R.: (1994) "Theory of Mind Is Contagious: You Catch It From Your Sibs", *Child Development 65*, pp. 1228-38

Perruchet, P. & Gallego, J.: (1997) "A subjective unit formation account of implicit learning", in: Berry, D.C. (1997), pp. 124-61

Pink, T. L. M.: (1991) "Purposive Intending", *Mind 100*, pp. 343-59

(1997) "Reason and Agency", *Proceedings of the Aristotelian Society XCVII*, pp. 263-80

Premack, D. & Woodruff, G.: (1978) "Does the chimpanzee have a theory of mind?", *The Behavioral and Brain Sciences 4*, pp. 515-526

Putnam, H.: (1973) "Explanation and reference", in Pearce, G. & Maynard, P.: (eds.) *Conceptual Change*, Dordrecht-Reidel, pp. 199-221, reprinted in his: *Mind, Language and Reality: Philosophical Papers Vol. 2*, Cambridge University Press, Cambridge, 1975

(1975a) "The meaning of 'meaning'", in Gunderson, K.: (ed.) *Language, Mind and Knowledge*, University of Minnesota Press, Minneapolis, MN, pp. 131-93

(1975b) "Philosophy and Our Mental Life", in his: *Mind, Language and Reality: Philosophical Papers vol. 2*, Cambridge University Press, Cambridge, pp. 291-303

Ramsey, F. P.: (1978) "Theories", in his *Foundations*, ed.: Mellor, D.H., Routledge & Kegan Paul, London, pp. 101-25

Ravenscroft, I.: (MS) "Predictive Failure"

Rey, G.: (1993) "Why presume analyses are on-line?", *Behavioral and Brain Sciences 16*, pp. 74-75

Ross, L. & Nisbett, R.: (1991) *The Person and the Situation: Perspectives of Social Psychology*, McGraw-Hill, New York, NY

Ryle, G.: (1949) *The Concept of Mind*, Hutchinson, London

St. John, M. F. & Shanks, D. R.: (1997) "Implicit learning from an information processing standpoint" in Berry, D.C. (1997), pp. 162-94

Searle, J. (1983) *Intentionality*, Cambridge University Press, Cambridge

Segal, G.: (1996) "The modularity of theory of mind", in: Carruthers, P. & Smith, P.K. (1996), pp. 141-57

Sellars, W.: (1963) "Empiricism and the Philosophy of Mind", reprinted in his: *Science, Perception and Reality*, Routledge & Kegan Paul, London, pp. 127-96

Semin, G. R. & Gergen, K. J.: (1990) (eds.) *Everyday Understanding*, Sage Publications Ltd., London

Scholl, B. J. & Leslie, A. M. Leslie (1999) "Modularity, Development and 'Theory of Mind', *Mind & Language 14*, pp. 131-53

Shoemaker, S.: (1968) "Self-Reference and Self-Awareness", *Journal of Philosophy 65*, pp. 555-67

(1990) "First-Person Access", in *Philosophical Perspectives 4*, pp.187-214

Smedslund, J.: (1990) "Psychology and Psychologic: Characterization of the Difference", in: Semin, G. R. & Gergen, K. J. (1990), pp. 45-63

(1997) *The Structure of Psychological Common Sense*, Lawrence Erlbaum Associates, Mahwah, NJ

Smith, E. R. & Miller, F. D.: (1978) "Limits on Perception of Cognitive Processes: A Reply to Nisbett and Wilson", *Psychological Review 85*, pp. 355-62

Smith, M.: (1994) *The Moral Problem*, Blackwell, Oxford

Smolin, L.: (1997) *The Life of the Cosmos*, Weidenfeld & Nicholson, London

Stein, E.: (1996) *Without Good Reason*, MIT Press, Cambridge, MA

Sterelny, K.: (1993) "Categories, categorisation and development: Introspective knowledge is no threat to functionalism", *Behavioral and Brain Sciences 16*, pp. 81-3

Stich, S.: (1978) "Beliefs and Subdoxastic States", *Philosophy of Science 45*, pp. 499-518

(1983) *From Folk Psychology to Cognitive Science*, MIT Press, Cambridge, MA

(1996) *Deconstructing the Mind*, Oxford University Press, Oxford

Stich, S. & Nichols, S.: (1992) "Folk Psychology: Simulation or Tacit Theory?", *Mind and Language* 7, pp. 35-71, reprinted in: Davies, M. & Stone, T. (1995a)

(1995) "Second Thoughts on Simulation", in: Davies, M. & Stone, T. (1995b), pp. 87-108

(1996) "How Do Minds Understand Other Minds? Mental Simulation versus Tacit Theory", in: Stich, S. (1996) *Deconstructing the Mind*, Oxford University Press, Oxford, pp. 136-67

(1997) "Cognitive Penetrability, Rationality and Restricted Simulation", *Mind and Language 12*, pp. 297-326

Stich, S. & Ravenscroft, I.: (1996) "What *Is* Folk Psychology?", in: Stich, S.: *Deconstructing the Mind*, Oxford University Press, Oxford, pp. 115-35

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

Tversky, A. & Kahneman, D.: (1993) "Probabilistic Reasoning", in: Goldman, A.: (ed.) *Readings in Philosophy and Cognitive Science*, MIT Press, Cambridge, MA, pp. 43-68

van Fraassen, B. C.: (1980) *The Scientific Image*, Clarendon Press, Oxford

Wason, P. C. & Johnson-Laird, P. N.: (1972) *The Psychology of Reasoning*, Batsford Ltd., London

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

White, P. A.: (1988) "Knowing more about what we can tell: 'Introspective access' and causal report accuracy 10 years later", *British Journal of Psychology 79*, pp. 13-45

Wilkes, K. V.: (1984) "Pragmatics in Science and Theory in Common Sense", *Inquiry 27*, pp. 339-61

Wilson, T. D.: (1985) "Strangers to Ourselves", in: Harvey, J. H. & Weary, G.: (eds.) *Attribution. Basic Issues and Applications*, Academic Press Inc., Orlando, FL, pp. 9-36

Wimmer, H., Hogrefe, J. & Sodian, B. (1988) "A second state in children's conception of mental life: Understanding informational accesses as origins of knowledge and belief", in: Astington, J.W., Harris, P.L. & Olson, D.R.: (eds.) *Developing Theories of Mind*, Cambridge University Press, Cambridge, pp. 173-92

1. p.13: FP = practice.

So how can generalisations like G1-G8 be the <u>core</u> of FP?

2. G1-G8 are supposed to be part of (internal) FP. How do we know that it is those particular generalisations which are internally represented, rather than some other (behaviourally equivalent) set?

3 In what sense of 'know' is it true that everyone <u>knows</u> G1-G8, knows that everyone knows, etc (pp35-6)? Not everyone can <u>articulate</u> them.

4. p.36 Claim of big differences in FP capability. Evidence?
- Claim must be due to competence --- argument?

5 p.48. What is "pretend-seeing"?

6 pp 48-9 Claims about ~~re~~ simulation

of relative studies etc ~~not~~ well supported

7. p.57. ST and explanation – no covering law.

8. p.61 and seems a bit muddled on the int/ext distinction.
elsewhere.

9. p.62 (see also SGN quote on p.56) ---- Misreading of SGN's ~~[illegible]~~ interpretation of "theory".

10. Around p.62 Slides b/w ~~wee~~ folk theory = inclusive sense (ie just a body of knowledge/belief)

and folk theory = inclusive – scientific.

11. pp. 66-7: metarep used to distinguish TT from ST; But Heal points out that, plausibly, the simulator must begin with states like 'T believes that p'.

12. pp 69-70. Nested simulation.

13. pp 70-1. Simulation of <u>decision</u> involves the decision-maker; simulation of <u>reasoning</u> involves the theoretical reasoner. Goldman made this point in (1989).

14 p.71. A simulation involving metareps is <u>not</u> indistinguishable from a TT case.

15. p.85. See bottom for note on ambiguity of 'theory'.

16. p.86. Need to say what these (alleged) horrible consequences are.

17. p.88. TT could claim that FP is sufficiently "scientific" to count as a theory, but deny that it could form the basis of a good scientific psychology: Churchland holds that view.

~~18. p.94. which very prominent scientific theories~~

18. p.96. Relativisation of 'observable'.

19. p.101. Need to say more about what a "lawlike relation" is. Are there 'lawlike relations' in the theory of natural selection? If so, what is it? (Historical sciences seem to form a special case.)

20. p.103. How does proposed rephrasing of (iv) avoid the indicative law problem?

21. p.103 — See comments on proposed account of ~~law~~ theory.

22. p.104 – Which sense of FP(I) or (E) is being used here? Claim only plausible for FP(E).

23. p.105 Proverbs = slogans. FP gets taken "at face value". Int or ext. No arg. for this claim.

24. pp105-6 "Abs. makes the heart grow fonder." Sounds causal/explan

25. p.109. cf Twin Earth intuition.

25. p.116 Specific v f/work theories: see comments on p.116.

~~26 p.126. Semantic functionalism & TT — int or ext?~~

26 p.139 Unclear what is supposed not to be given directly. Content of the state or that it has content?

27. pp140-2 Discussion of R. – Not quite right. R offers an account of why we go right, and ~~then~~ endorses a 'mostly false of' line on why we say what we say.

28. pp140-2. S picks right. Applies theory. Says 'They were the best'. R fails to predict (R) – by applying the theory? Asymmetry.

29. pp148-9: Direct access conferral: how do we do it?

30. p.154. Why is semantic functionalism 'built into' classical & characteristic functional theories of self-attribution?

31. p. 156. Why isn't conceding direct access a retreat from TT?

32. p. 156. Sellars as first TT? Not really – Herder & ~~other~~ others.

p. 199-200 Smedslund's list of (1) – (13) is rather unconvincing – most of them seem either false or controversial.

1. <u>FP</u>

1. What is the status of G1 – G8.

- Int or Ext? Look like Ext.
- By analogy with grammar, Int may not be all that recognisable.
- In what sense does everyone know G1-8. pp55-6?
  Can't articulate them.
- Swedlund's list looks looks implausible. Most seem either wrong or controversial.
- qualia of recognition: qualia of 'yes, that's what I've always believed' or qualia of 'yes, I see it must be like that'?
- FP inferentially integrated? cf syntactic processing. Where ~~does~~ is contextual info brought to bear?

~~2.~~ "~~Theory~~".

2. ST/TT.

Metarep. to distinguish ST from TT: Heal's point. Simulation starts with beliefs with contents like "T believes p/desires q."

3. Theory.

Account of p.103. "lawlike" — theory of nat. selection?
"refer to entities" — realism?
coherent/interrelated?

4. Position Effect.

- FP is used to ~~[illegible]~~ give an account of why the subject said 'I chose the best'.
  FP is also used to explain ~~how~~ how we predict what others will do ~~—~~
  So why the asymmetry?
- R. is ~~[illegible]~~ misrepresented. ~~[illegible]~~ It's an account of how a simulationist can explain-pred. failure in this case. Good point. ST has the same problem with asymmetry as does TT.