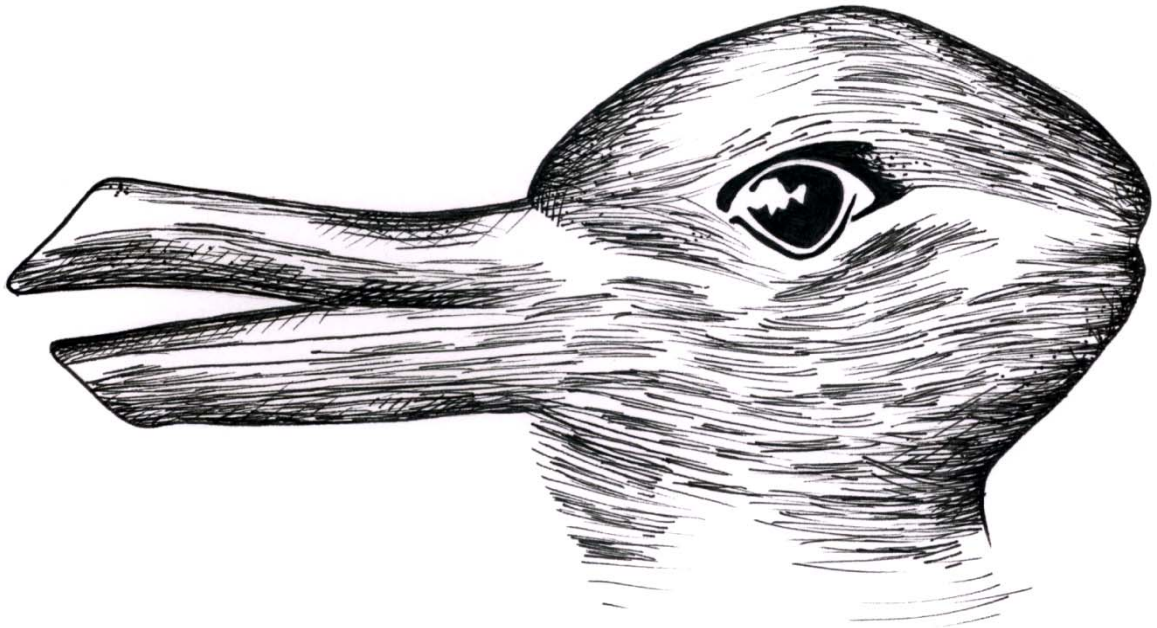


THEORETICAL SELF-INTERPRETATION



Andrew Garford Moore

2009

ABSTRACT

This essay aims to answer the following question: How do adult Homo sapiens self-ascribe intentional states. In other words how can a human agent come to know that he currently believes that x, intends that y or desires that z etc; where x, y and z are objects or states of affairs in what Bertrand Russell (1996 [1914]) referred to as the “external world”? The orthodox position is that we have direct and reliable introspective access to our intentional states. I argue that the introspective account is inaccurate and develop an alternative account in which self-ascription of intentional states is achieved by a process of theoretical self-interpretation. This line of argument was developed by Wilfrid Sellars (1956) in order to counter what he called the myth of the given, which in this context is the idea that knowledge of our own intentional states is “given” to us by a direct, infallible mechanism. Sellars invented the “myth of Jones”, an account of ascription in which intentional states are the posits of a theory of mind (TOM) and are ascribed to others interpretively on the basis of behavioural observation. This mechanism is then turned upon ourselves enabling the self-ascription of intentional states. By developing an alternative mechanism Sellars hoped to remove support for the dominant introspective position. Sellars’ myth came in for heavy criticism and proved unpersuasive. In this essay I aim to provide a more defensible version of the theoretical self-interpretation account by utilising advances in cognitive science and the philosophy of science. In the critical chapter 1 I aim to remove support for the introspective account by arguing that the introspective intuition provides no support for an introspective account of self-ascription. I then examine evidence showing that at least some of the time self-ascription is achieved by a process of self-interpretation. I conclude from this chapter that if a plausible self-interpretive account of the mechanism of self-ascription can explain all instances of self-ascription then there is no reason to posit a separate introspective mechanism. In the remaining two chapters I aim to provide such an account. In the chapter 2 I demonstrate that intentional states are the implicitly defined posits of a TOM. I delineate 4 components of a theory and after drawing a distinction between knowledge and

information examine the relation of these components to ascription. The third component will be identified with the TOM ascription mechanism. In chapter 3 I show how the possession of a TOM enables the theory-laden perception of intentional states, explaining the phenomenological directness of self-ascription. I go on to describe the type of information available to the TOM ascription mechanism and how this enables us to self-ascribe even when we exhibit no overt behaviour. I conclude that intuition provides no support introspection. Further to this theoretical self-interpretation has greater explanatory power than introspection as it is able to explain instances of self-ascription that introspective theory cannot. In order to achieve the same explanatory power the introspectionist resorts to a hybrid 'dual method theory' (e.g. Goldman 2006) and this is rejected for reasons of parsimony. Therefore introspective self-ascription should be rejected in favour of a purely theoretical self-interpretation account.

Cover image by Gavin Wells. It is a version of the ambiguous duck/rabbit originally found in Jastrow (1899). The same visual information is interpreted in two different ways. The image suggests that what we consciously perceive is mediated by concepts and not raw sense information.

CONTENTS

INTRODUCTION	6
1 THE INTROSPECTIVE INTUITION	12
1.1 Split brains and the introspective intuition	13
1.2 Self-interpretation in healthy adult subjects	15
2 INTENTIONAL STATES AS THEORETICAL POSITS	19
2.1 Folk psychological abilities exploit a body of information	20
2.2 The theoretical nature of the body of information subserving folk psychological abilities	25
2.21 Mendel's theory of inheritance – A paradigm of a scientific theory	25
2.22 The positing of unobservables and implicit definition of terms	27
2.23 Prediction by projection	29
2.24 Deep causal explanation	30
2.25 Cognitive economy	30
2.26 Interpretation	32
2.3 The unenunciability objection and the erroneous appeal to tacit knowledge	34
2.31 Tacit information and conceptual knowledge	35
2.32 Conceptual knowledge of TOM	37

2.4 Theories as models and hypotheses – A response to the unenunciability	
Objection	39
2.4.1 Scientific theories as models and hypotheses	39
2.4.2 TOM as models and hypotheses	43
2.5 Components of theories and ascription	45
3 THEORETICAL INTERPRETATION	48
3.1 Theory-laden perception	48
3.2 What is observed when we perceive intentional states?	54
CONCLUSION	58
BIBLIOGRAPHY	60

INTRODUCTION

It is uncontroversial to say that humans have an ability to explain and predict the actions, decisions or inferences of themselves and other human beings. These abilities are known as *folk psychological abilities*. For example as an explanation for why Oliver did X we could offer the following:

Oliver did X because he *desired* Y and *believed* that doing X would bring it about that Y

The same resources can also be used in prediction, suppose Oliver had not yet done X, we could predict that he would indeed do X, given the information that he *desires* Y and that he *believes* that doing X would bring it about that Y.

In the above example we are ascribing certain mental states to the human subject, I will refer to these initial ascriptions as input. In the above example the input is Oliver's *desire* for Y and his *belief* that performing X will bring it about that Y. These mental states are assigned a causal role in the production of what I shall refer to as *output*; either a decision, an action or an inference (Saxe 2005 p174). By ascribing certain inputs we are able to predict or explain the occurrence of a particular output. In the above example the output is the behavioural prediction that Oliver will perform X. Depending on the output we can arrive at either a behavioural prediction or the ascription of a further mental state.

In order to explain or predict Oliver's output we ascribe to him the mental states *belief* and *desire*.

These mental states are examples of a class of mental states known as intentional states. An intentional state expresses an agent's attitude to a proposition and possesses a characteristic aboutness or reference to objects or states of affairs in the external world. Take the proposition P, it is possible for an agent to have many different attitudes towards this proposition, from the above examples we can see that Oliver can *believe* that P and he can *desire* that P, but he can also *hope* that P, *intend* that P and *fear* that P and so on. In this essay when I use the term "mental state" it

should be taken to refer exclusively to the subclass of mental states known intentional states rather than the broader definition of mental states, which includes such things as perceptions and emotions.

Not only do we ascribe mental states to others to predict and explain their output, we also ascribe them to ourselves. The knowledge that we *believe* that X, *intend* that Y or *desire* that Z seems to just come to us. The process by which we acquire knowledge of our own occurrent intentional states is the process of *self-ascription*. My primary aim in this essay is epistemological, to provide a positive account of self-ascription of intentional states to be used as **input** for folk psychological inferences.

The dominant account of self-ascription is introspection. Individual accounts differ in detail but what they have in common is the idea that we have direct and reliable knowledge of our own intentional states. In contrast, as we cannot have direct access to the mental states of others, other-ascription must proceed by some kind of interpretation of behaviour (including verbal behaviour).

Introspection therefore carries the commitment that the mechanism of self-ascription is different in *kind* to the mechanism of other-ascription.

On the face of it introspection certainly seems the most plausible account. Nothing seems as certain as the knowledge we have of our own mental states. And we intuitively believe that our knowledge of ourselves is different to knowledge we have of others. In this essay I argue that the introspective intuition is misleading and that we self-ascribe by a process of *theoretical self-interpretation* using the same mechanism as other-ascription. That is we self-ascribe in the same way as we other-ascribe: by observing and integrating a wide range of evidence to form an ascription judgement. This evidence includes such things as overt behaviour (including verbal behaviour), current circumstances and memories of past observations. I contend that our knowledge of our intentional states is not given, it needs to be extracted.

The theoretical interpretation view defended here originated in the highly original philosophy of Wilfrid Sellars and is set out in his 1956 essay 'Empiricism and the Philosophy of mind'.

Sellars main concern was to attack the idea of *the given* supporting foundationalist philosophies, which in the context of the philosophy of mind is the idea that we have direct introspective access to our own intentional states. In order to raise doubts about the dominant introspective account of self-ascription, Sellars invented a myth¹. The function of the myth was to demonstrate a way in which we could gain knowledge of our intentional states without committing ourselves to the idea that some knowledge is given to us directly and infallibly. By providing a possible alternative Sellars hoped to remove what he saw as a major attraction of foundationalist philosophies, the belief that they are the only plausible way of explaining human knowledge.

The essential features of Sellars' myth are (a) the theoretical nature of intentional states, he argued that intentional states were 'in' the mind in the same way that molecules are 'in' the air around us. That is, what intentional states *are* is determined by a TOM. (b) That the same interpretive mechanism that is used to infer the states of others is used to infer one's own intentional states.

That Sellars' account has not proved persuasive is evidenced by the predominance of the introspective account. The reasons for this are threefold.

(1) We have a strong intuition that knowledge of our own intentional states is direct and infallible, giving plausibility to the introspective account of self-ascription and making self-interpretation seem implausible.

(2) The unenunciability objection to the TOM theory (if we have a TOM why are we not able to state its generalisations?) raises doubts about Sellars' claim that intentional states are theoretical states.

(3) The belief that even if TOM theory was consistent, Theoretical self-interpretation theory is unable to: (3.1) account for the phenomenological directness of intentional state knowledge. Or

¹ Sellars' myth of Jones is a fictional account of how our ancestors gained intentional state concepts.

(3.2) explain instances of self-ascription in the absence of overt behaviour (the evidence on which other-ascription is based).

In this essay I intend to make up these deficiencies. Reviewing recent advances in cognitive science (Brasil-Neto *et al* 1992 & Gazzaniga 1995) I will show that the introspective intuition provides no support for the introspective account and further that some instances of self-ascription can only be made by self-interpreting. Then, by reworking TOM theory using advances in the philosophy of science (Giere 1988, 1999 & Maibom 2003) I develop a version of TOM theory that is consistent with our inability to state theoretical generalisations. Finally I will describe how theory-laden perception accounts for the phenomenological directness of intentional state knowledge. And that access to a variety of different information sources (Carruthers 2009) allows the theoretical self-interpretation account to explain *all* instances of self ascription. This will be achieved as follows:

In chapter **1** I demonstrate that our intuitions regarding self-ascription are misleading and that a purely introspective account is not consistent with the evidence. In section **1.1** I demonstrate that the intuitive plausibility of introspection is misleading by describing a study in which commissurotomy subjects self-interpret but retain the introspective intuition. In section **1.2** I describe a second study in which healthy adults are shown to self-interpret, demonstrating that - contrary to intuition - at least some self-ascription is achieved by self-interpretation. The remainder of the essay consists of the development of a positive account of theoretical self-interpretation.

In chapter **2** I demonstrate that folk psychological abilities are subserved by a theory of mind (TOM) and that this is consistent with our inability to state theoretical generalisations. I delineate 4 components of theories the third of which is identified as the TOM ascription mechanism. In section **2.1** I establish that folk psychological abilities are subserved by a body of information concerning the interaction between intentional states. In section **2.2** I show that the body of information subserving folk-psychological abilities is structured theoretically. I delineate a provisional list of 3 theoretical

components. In section **2.3** I raise the uneuncibility objection to TOM theory and demonstrate that the traditional TOM theorists appeal to a tacitly known theory is inadequate. I achieve this by clarifying the tacit information/conceptual knowledge distinction and demonstrating how the two components that need to be construed as tacit information to avoid the objection are in fact part of our conceptual knowledge. In Section **2.4** In order to rebuff the uneuncibility objection I show that theories are structured as a collection of models and hypotheses rather than generalisations as the objection assumes. In section **2.5** I finalise the list of theoretical components by adding a fourth and examine the role of each in theoretical prediction/explanation. I identify the third theoretical component with the TOM ascription mechanism enabling both other and self-ascription.

With the theoretical nature of intentional states established, chapter **3** shows how the TOM self ascription mechanism is capable of explaining the phenomenological directness of intentional state knowledge and instances of self-ascription in the absence of overt behaviour. In section **3.1** In order to explain the phenomenological directness of intentional state knowledge, I draw a distinction between *observation* and *perception* and describe how a body of domain specific information allows direct theory-laden perception of unobservable theoretical posits. In section **3.2** I explain that by *observing* information such as overt behaviour, visual imagery and inner speech and integrating it with stored information; the TOM ascription mechanism is able to self-ascribe even in the absence of overt behaviour.

I conclude that the version of theoretical self-interpretation theory developed here has greater explanatory power than a purely introspective account and the interpretive account should be favoured. Further to this, hybrid dual method theories that retain introspection are to be rejected for reasons of parsimony. Therefore introspection theory should be abandoned in favour of a theoretical self-interpretation account. The argument however is not decisive and I suggest that the final answer to the question will be decided empirically.

Before embarking on the above task, I will use the remainder of this introduction to justify the use I intend to make of empirical data in my epistemological investigations. I will take for granted what W. V. Quine (2004 [1951]) argued for in his 'Two Dogmas of Empiricism': that there is no way in principle to draw a sharp distinction between analytic and synthetic truths and hence philosophy does not work with a class of truths that are wholly insulated from observational knowledge. The consequence of this distinction blurring for epistemology is that knowledge cannot be "rationally reconstructed" (Quine 2004 [1969]) from an infallible base, the best we can do is to understand and make clear the (potentially fallible) connection between sensory input (or possibly quasi sensory input of the introspective variety) and human knowledge. In elucidating this connection, I will therefore occasionally make use of empirical knowledge - mainly from neuroscience and psychology - to support my philosophical conclusions. It could be objected that the appeal to scientific knowledge is circular, that in clarifying the foundations of human knowledge, it is not permissible to appeal to a subset of that knowledge (scientific knowledge). However, if rational reconstruction is indeed impossible, if it impossible to *deduce* knowledge from sensory or quasi sensory input and our task is rather to simply understand the connection between sensory input and knowledge "we are well advised to use any available information, including that provided by the very science whose link with observation we are seeking to understand." (Quine 2004 [1969] p264). The naturalistic approach sketched above construes philosophy as a subset of - and subservient to - science, so what role has the naturalistic philosopher left to play in the search for knowledge? As Carruthers (2006) points out "science always contains what might be called 'framework assumptions', as well as detailed theories closely grounded in the empirical data. And the examination and defence of those assumptions can be the work of naturalistically minded philosophers". In other words the task of the naturalistically inclined philosopher is to probe, clarify and occasionally criticise the foundations on which the edifice of human knowledge constructed.

1

THE INTROSPECTIVE INTUITION

When we ascribe intentional states to ourselves it certainly does not feel that we arrive at such knowledge through a process of self-interpretation, we *just know* that we believe X, desire Y or intend that Z. Intuition therefore seems to support some kind of introspective account and makes self-interpretation seem implausible. The history of science and philosophy however shows that intuitions are not always a reliable method of arriving at veridical conclusions. In this chapter I will demonstrate that our intuitions regarding intentional state knowledge are misleading.

Many recent authors embrace a direct monitoring theory of access to current intentional states (e.g. Goldman 2006 and Nichols & Stich 2003). Direct monitoring theories differ in detail, but essentially they claim that there is a direct link between the occurrence of an intentional state and the agent having knowledge of that event. They are therefore committed to the claim that the mechanism of self-ascription is different in kind to that of other-ascription. The mechanism of self ascription posited by such direct access theories can be seen as a species of *introspection*. Carruther's (2009 p123) broadly defines introspection as "any method for forming beliefs about one's own mental states that *is not* self-interpretive and that differs in *kind* from the ways that we form beliefs about the mental states of others".

I intend to show firstly that the introspective intuition provides no support for an introspective mechanism and secondly that despite its intuitive implausibility, self-interpretation is a viable method of self-ascription.

In section 1.1 I will describe a study performed with split brain subjects by the cognitive scientist Michael Gazzaniga (1995) in which subjects self-ascribe interpretively with all the confidence of normal everyday self ascription, thus demonstrating that the introspective intuition does not count

in favour of a separate introspective mechanism. In section **1.2** I will describe a study which shows that it is not only split brain subjects that self-interpret but normal healthy adults as well. This demonstrates that despite the intuitive implausibility of self-interpretation, it is a viable method of self-ascription.

Intuition therefore provides no support for an introspective mechanism and does not count against a self-interpretive account, which is shown to be a viable method of self-ascription. To explain examples of self-interpretation, the introspection theorist is therefore forced to defend a hybrid account.

1.1 Split brains and the introspective intuition

This section will demonstrate that the intuitive plausibility of introspection is misleading by describing a study in which commissurotomy subjects self-interpret but retain the introspective intuition. The human brain is divided into two separate hemispheres. Inter-hemispheric communication is achieved by means of a thick bundle of nerve fibres known as the *corpus callosum*. Certain symptoms of severe epilepsy can be reduced by a commissurotomy, a surgical procedure which involves the severing of the corpus callosum, effectively preventing inter-hemispheric communication. Surprisingly, patients that have undergone this procedure report little difference in their sense of their self as a unified being (Gazzaniga, 2000 p 1309)². Over the past 40 years, patients who have undergone a commissurotomy have been studied extensively by Gazzaniga and others. The results of their studies are a valuable resource for researchers from a variety of fields interested in the human mind. Below I shall describe one of Gazzaniga's more recent studies and then show how it undermines the introspective intuition.

The study (Gazzaniga, 1995) involved the presentation of different visual stimuli to each individual brain hemisphere. With the subject looking straight ahead, stimuli could be presented exclusively to

² This is not surprising however – for reasons that will emerge later - if one accepts the self-interpretive position espoused in this essay.

the right eye and thus the left hemisphere or to the left eye and thus the right hemisphere.

Language production is performed by the left hemisphere but both hemispheres are capable of understanding language. The subject's verbal reports therefore come exclusively from the left hemisphere. When a card with the instruction "walk" was presented to the left eye (the right hemisphere) the patient stood up from the table and went to exit the room. When the tester asked why this was, the patient (or rather his left hemisphere) replied "I'm going to get a coke from the house".

Upon being asked to explain his behaviour, the subject self-ascribes an intentional state. In this case the intentional state in question is the *intention* to get a coke. But clearly this self-ascribed intention is not what caused him to get up from the table. The subject's left hemisphere reporting the intention has no access to the real cause of his behaviour and falsely ascribes an intention in order to explain the behaviour. It seems reasonable to conclude that the subject's self-ascription was achieved by a process of self-interpretation, for such an error cannot be accounted for by introspection. Perhaps the subject felt thirsty (access to non-intentional feelings with attendant qualia such as thirst is not denied in this essay) and this feeling taken in conjunction with the subject's taste for coke and the observed behaviour of leaving the room led him to (falsely) self-ascribe the intention to get a coke. The important point for the present argument is that this false self-ascription is delivered with all the speed and confidence of a normal everyday self-ascription. The subject showed no awareness of self-interpreting and still demonstrated the intuition that he was introspecting. This point is supported by the fact (reported by Carruthers (2009) p126 from a personal communication from Gazzaniga) that during testing the patient was reminded by the tester of his operation and how it can affect access to what is seen in his left visual field. Testing was stopped at several points during the process and the tester reminded the subject "Joe, as you know, you have had this operation that sometimes will make it difficult for you to say what we show you over here to the left of fixation." As Carruthers (2009) notes these reminders would be expected to make the subject aware that some of his behaviour may be caused by what is shown to the right

hemisphere, but that was not the case. After being reminded of his operation the subject continued to falsely self-ascribe intentional states to explain his behaviour. "If patients were aware of interpreting rather than introspecting, then one would expect that a reminder of the effects of a commissurotomy would enrich the hypothesis pool, and would sometimes lead them to attribute some of their behaviour to that. But it doesn't do so." (Carruthers, 2009 p126)

The above study demonstrates that the introspective intuition persists despite the subject demonstrably self-ascribing an intentional state by self-interpreting. From this we can conclude that the introspective intuition provides no support for introspection of intentional states.

1.2 Self-interpretation in healthy adult subjects

Although Gazzaniga's studies show that the introspective intuition provides no support for the introspection of intentional states, we cannot conclude from them that healthy adults also self-interpret. This is because the studies involved subjects that had undergone a rare surgical procedure. It may be that self-interpretation is limited exclusively to individuals who have undergone this procedure. In this section I will show that this is not the case by describing a study in which healthy adult subjects can be shown to self-ascribe by self-interpretation.

A study by Brasil-Neto *et al* (1992) involved magnetic stimulation of the subject's motor cortex. Subjects were asked to extend either their right or their left index finger after hearing a click. They were asked to decide which finger to extend only after hearing the noise. What the patients were unaware of was that the magnetic stimulation of the motor cortex caused a bias in their choice of finger. Subjects tended (to a statistically significant degree) to extend the finger contralateral to the brain hemisphere being stimulated compared with trials involving frontal cortex stimulation and with trials involving no stimulation. Effectively, which finger was extended was influenced by the magnetic stimulation. But importantly for my argument, subjects were unaware that it was the magnetic stimulation that caused this or that finger to extend and continued to believe that they

were making a truly free *decision*. In other words, they ascribe to themselves an intentional state - namely the *decision* to extend this or that finger - in order to explain their behaviour. But this is plainly false, since the extension of the finger was determined by the magnetic stimulation. This false ascription provides strong support for the self-interpretive account. The intentional state which is intuitively the cause of the finger extension is shown to be itself caused by the finger extension process.

It could be objected that the magnetic stimulation itself caused the introspectible decision which in turn caused the finger extension. But – as Carruthers (2009) notes - if what the magnetic stimulation caused in the motor cortex was a *decision* event, for it to be describable as introspectible it must directly cause conscious knowledge of that event. In order for this to happen the back projecting pathways linking the motor cortex with the frontal cortex would need to carry the information of the occurrence of the decision back upstream from the motor cortex to activate the relevant intentional state concept. This process could be accurately described as direct monitoring or special access. However, if the back projecting pathways were used for this purpose we would expect stimulation of pre motor cortex to achieve the same result as motor cortex stimulation, yet stimulation of this area produces no choice bias, suggesting that the back projecting pathways do not carry this information. If this is the case, then the only possible route for the information of the occurrence of the *decision* event back upstream is through the perception of behaviour, most directly through the proprioception system. This would be a clear example of self-interpretation.

It is clear from the studies discussed above that any account that posits direct monitoring as the exclusive method of intentional state ascription is not consistent with the increasingly large body of scientific work on false self-ascription. The study in **1.2** showed that healthy adults can falsely ascribe intentional states to themselves and therefore must be self-interpreting. The example is significant, for it represents an instance of self-ascription that cannot be explained by introspective theory. In order to take into account this and other examples of self-interpretation, introspective

theorists need resort to a hybrid account, or what Goldman (2006 p232) calls 'dual method theory'. The theory allows that occasionally humans may self-ascribe by self-interpretation, but maintains the claim that the primary method of self-ascription is introspective. As Goldman (2009 p232) notes, although self-interpretation can be demonstrated "It does not follow, however, that confabulation is the only method, or the principal method, of first-person attribution". This is true, but in light of the study described in 1.1 showing that the introspective intuition provides no evidence of introspection, the concession leaves the introspectionist in a difficult position. Self-ascription by self-interpretation has been demonstrated scientifically while introspection of intentional states hasn't. With the support of intuition removed, what reason is there to suppose that introspection of intentional states occurs at all? If all self-ascription can be explained in terms of self-interpretation, then there is no work left for introspection to do. Even though its existence cannot be directly shown, Goldman (2006 p230) argues "That people use something like introspection can be made compelling by considering the implausibility of the alternatives. I believe that I currently intend to walk into my study and remove a particular book from the shelf. What leads me to think that I have this intention?"

This line of reasoning was what Sellars aimed to deter with his myth of Jones. But he underestimated the myth against which he was arguing³. At the time there seemed no reason to doubt the introspective intuition and Sellars account of the theoretical nature of intentional states came under criticism⁴. Consequently, introspection has remained the dominant account of self ascription. However, in the past fifty years much has changed. The evidence from cognitive science reviewed in this chapter has given us reason to doubt the introspective intuition and has demonstrated that we sometimes self-interpret. Further to this, advances from the philosophy of science have shown us

³ That is the myth of the given which Sellars saw in the dominant Cartesian position of the time. A reason for the intuitive plausibility is suggested by Carruthers (2008). He provides strong evolutionary arguments in favour of the innateness of introspective intuition. He argues that a Cartesian model of the mind would be able to form more accurate beliefs and also be far more efficient in terms of information processing compared to one that took self-interpretation into account.

⁴ Most menacingly, the enuncibility objection dealt with in sections 2.3 & 2.4

how objections to TOM miss the point. The following chapters aim to exploit these advances to remove the final prop of support from the introspectionist position: the lack of a convincing alternative. This will be achieved by building a plausible positive account of theoretical self-interpretation that is able to explain all instances of self-ascription. With this account in place and defended from objections, we can finally put introspection of intentional states to rest, over 400 years after it was articulated by Descartes and 50 years after Sellars showed what an alternative could look like.

2

INTENTIONAL STATES AS THEORETICAL POSITS

The first task of developing an account of self-ascription is to examine the nature of intentional states themselves. In this chapter I argue that folk psychological abilities are subserved by a TOM and that intentional states are the implicitly defined posits of this TOM. With this established it will allow me in chapter 3 to demonstrate how a component common to all theories enables self-ascription through a process of theoretical self-interpretation. TOM theory however has come in for much criticism in the years since it was suggested by Sellars. To ensure that my argument is built on solid foundations it is therefore necessary to describe in detail the structure of TOM to ensure that it is adequately defended from objections.

In section 2.1 I will show that folk psychological abilities rely on a body of information that describes the causal relations between intentional states. This will be achieved by providing objections which undermine the only viable alternative this position: Simulation theory. In section 2.2 I will show that the body of information that subserves folk psychological abilities has a theoretical structure. This will be achieved by describing a paradigm example of a scientific theory (Mendel's theory of inheritance) and highlighting the essential features of it. The body of information subserving folk psychological abilities will then be shown to share these essential features. It will therefore be concluded that folk psychology abilities are subserved by a TOM. It will be shown that intentional states are posits of TOM and that intentional states gain their meaning in the same way as posits of scientific theories, that is, through *implicit functional definition*. It will also be noted in this section that in order to gain definition in this way it is necessary that knowledge of TOM must be conscious. In section 2.3 I will raise and respond to a possible objection to TOM theory: The unenunciability objection. Essentially the objection runs as follows: If knowledge of TOM was conscious then we

would be able to state law-like generalisations which describe the causal interactions between intentional states. Since we are unable to state such generalisations TOM is therefore false. The traditional response to this objection - the appeal to a tacitly known theory - will be shown to be unsatisfactory. A provisional list of three theoretical components will be delineated. **2.4** In order to successfully counter this objection I will show that TOM is structured as a collection of theoretical models and hypotheses, not as a set of law like generalisations. So while we can say what we know, the answers are not structured in the way that the objector demands as any generalisation we do state is likely to be false. The unenunciability objection will therefore be rejected. **2.5** Adds a fourth theoretical component and discusses their relation to ascription. The third component will be identified with the ascription mechanism.

2.1 Folk Psychological abilities exploit a body of information

My task in the first section of this chapter is to establish the nature of the process that subserves folk psychological abilities. Accounts of this process can generally be classified into one of two categories: Information rich processes and information poor processes (Nichols & Stich 2003). I will show that folk psychological capacities are subserved by information rich process by providing objections to the alternative: information poor processes. I shall achieve this by describing an example of folk psychological error that cannot be explained by simulation theory. I conclude that Folk psychology is subserved by a body of information. Before describing the example, it is necessary to clarify the information rich/information poor distinction.

Information rich accounts (e.g. Carruthers 2009) posit a set of conceptual representations⁵ of mental states and a body of information that describes the causal interactions of these mental states with environmental stimuli, each other and with the resulting output. At this stage, how this body of

⁵ Or rather metarepresentational states, that is, representations of representational states.

information is structured and/or encoded is not important. What is important is that on this account folk psychological abilities such as prediction and explanation are subserved by a body of information that is itself *not* implicated in the production of actual output. On the information rich account I am able to predict that Oliver would do Y because I have a body of information that tells me that when people *desire* X and *believe* that performing Y would achieve X, they will perform Y. Similarly I use the same body of information to predict and explain my own output. TOM theory is a particular species of the genus of information rich accounts in which the further claim is made that the body of information subserving folk psychological abilities is structured in way analogous to a scientific theory.

Information poor accounts (e.g. Goldman 2006) state that folk psychological abilities are achieved without the need for a rich body of information linking representational states, but are rather achieved by using the *same system* that actually causes the action, decision or inference output being predicted or explained. Output is caused by a practical reasoning system that takes environmental stimuli and transforms them into an appropriate behaviour. In order to predict the behaviour of myself or a third person, I simply take this system offline - that is decouple it from actual environmental input and output - and provide it with hypothetical inputs, the resulting output is then taken as an output prediction. On this account I am able to predict that Oliver would do Y because I input the hypothetical *desire* for X and the hypothetical *belief* that performing Y would achieve X into my practical reasoning system and take the resulting output – perform Y - as a behavioural prediction. Such accounts are known as **Simulation** accounts.

Before deciding which of these accounts is the most plausible, a little clarification is necessary due to certain unwanted implications of the term simulation. Take a computer simulation of a natural phenomenon such as the effects of an earthquake on a tall building. The computer simulation is subserved by a body of information that states how the force of the earthquake is transferred into

the building and information encoding the strength and flexibility of the materials used in its construction. Using this information the simulation is able to calculate the stresses on different parts of the building and predict how it would fare in the event of an earthquake of a specified magnitude. While this is correctly called a computer *simulation* it is clearly an information rich process as the prediction is based on the interaction of representations of the various entities involved in the simulation. In the same way it would be correct to say that we are simulating a mind when we use metarepresentations of mental states and information that specifies their interaction to form a prediction. In this essay the term simulation will refer solely to information poor accounts, whose core claim is that we use the *same system* that is used for practical reasoning to form an output prediction. In our earthquake example third person prediction/explanation would be analogous to shaking a scale model of the target building in a way similar to that caused by an earthquake to see how it reacts in such circumstances. Self prediction/explanation would be analogous to shaking the actual building. The analogy falls down in that in the earthquake simulation we are unable to decouple the result from reality in a way analogous to the simulation theorist claims we can decouple our practical reasoning system from its resulting output.

In order to argue against information poor accounts I will describe a psychological experiment that demonstrates that humans can often make systematic errors when predicting the behaviour of themselves and others. While these errors can easily be accounted for by information rich theories they are unable to be accounted for by information poor processes. Think again of the earthquake simulation⁶, if the prediction proved to be inaccurate, what could account for such inaccuracies? On the information rich account the inaccuracy could have occurred because of inaccuracies in the body of information that we use to arrive at the prediction, for example the values we use for the strength of the construction materials or how the earthquake's force transfers to the superstructure.

⁶ Ignoring for sake of argument the impracticalities of an information poor simulation of such an event.

On the information poor account however there are only two possible sources of error. Firstly error could occur because the model being used is in some way different to the target model. Secondly error could occur if we have not shook the building in the same way it would be shook by a by an earthquake of a particular magnitude, in other words because of an inaccurate input to the system.

Therefore if we can describe a psychological experiment in which subjects make inaccurate predictions of human behaviour, in which there is no possibility of an inaccurate input, and with a large enough sample size to eliminate the influence of differences between individual minds, we will be able to conclude that folk psychology is subserved by an information rich process.

There are many such experiments to be found in the psychological literature (e.g. Milgram, S 1963; Tversky & Kahneman 1974) but due to space constraints I shall concentrate on just one. Nichols and Stich (1996) describe an experiment by Ellen Langer (1975) in which subjects make systematic errors in their predictions of others behaviour. Langer arranged a football pool a few days before the Superbowl and sold tickets to workers in an office for one dollar each. Half of the subjects were able to choose their ticket and the other half were given no choice of ticket. A few days later on the eve of the game Langer offered to buy back the tickets and asked the subjects what they thought would be a fair price. The result was that those who had been offered a choice of tickets quoted a statistically significant higher price for their tickets than the no choice group. This surprising result has been dubbed by Nichols and Stich 'the Langer effect'.

A description of the experiment was read to a group of students who were then asked to predict the buy back price quoted by the choice and no choice groups. The students predicted no statistically significant difference between the buy back price quoted by the two groups. After criticism, the results were repeated in a similar experiment where the offering of tickets was captured on video which was then shown to the group of students predicting the buy back price.

These results are significant because if the predictions were made using an information poor process then the errors must have been due to one of the following factors (from Nichols and Stich 1996)

(a) The predictors decision making (or practical reasoning) mechanism is different from the targets

(b) The pretend belief and desire generator has not provided the decision making system with the right beliefs and desires – i.e. with the ones that actually motivate the target person whose behaviour is to be predicted.

The predictive error observed in this experiment could not be due to (a) assuming a random distribution in differences of practical reasoning systems, individual differences could not have caused a statistically significant difference between the actual price quoted and the predicted price. The error could also not be due to (b) because in the repeated experiment a video of the choice was shown so that all relevant inputs would have been as visible to the predicting group as they were to the target group. The experiment also suggests that the predicting group would not be able to predict their *own* behaviour if they were in the experimental situation.

While these results are unable to be explained by information poor process accounts, they are easily explained by information rich accounts. We would not expect an information rich process to model every intricacy and unexpected effect that we find in the actual mind, the error can simply be put down to an inaccuracy in the body of information that subserves such predictions. The inability of information poor process accounts to explain the results of this and similar experiments means that they must be rejected in favour of information rich process accounts that are able to explain such results. The conclusion of this section is therefore that folk psychological abilities are subserved by information rich processes, that is they depend on a body of information containing the possible

causal interactions of mental states with environmental input, each other and with the resulting output . The task of the next section is to determine how this body of information is structured.

2.2 The theoretical nature of the body of information subserving folk psychological abilities

My task in this section is to establish the theoretical nature of the body of information that is exploited when we use folk psychological abilities. Philosophers of science are yet to agree on how to precisely define theory. There are however several features that it is reasonably uncontroversial to say are unique to theory. In this section I will show that the body of information subserving folk psychological abilities possesses these unique features and therefore conclude that folk psychology is subserved by a theory of mind. Firstly in **2.21** I will describe a paradigm example of a scientific theory: Mendel's theory of inheritance and contrast it with pre Mendelian gardening lore, an example of a body of non theoretical empirical knowledge. In the sections following I will pick out what I contend are features unique to theories and show how they are shared by the body of information subserving folk psychological abilities. In **2.22** I shall discuss the positing of unobservable entities and their implicit functional definition, in section **2.23** Prediction by projection, section **2.24** Deep causal explanation, section **2.25** Cognitive economy and section **2.26** Interpretation. From my discussion of these features I will also delineate a provisional list of three functional components of TOM.

2.21 Mendel's theory of inheritance – A paradigm of a scientific theory

In the 19th century the scientist Gregor Mendel began experiments involving the cross pollination of different varieties of peas. One of the varieties of peas had purple flowers and the second variety had white flowers. The dominant theory of inheritance at the time was that of blending, the theory claimed that the characteristics of an offspring of any two organisms would be a blend of the characteristics of the parents. If this theory was correct it would be expected that the offspring of a

purple flowered pea and a white flowered pea would have flowers of a light shade of purple. What Mendel observed was that the second generation of offspring all had purple flowers of the same shade as that of the purple flowered parent. Continuing the experiment to the next generation Mendel found that cross pollination of the second generation of purple flowered peas produced offspring of either purple or white flowers at a ratio of 3:1. The blending theory of inheritance was unable to account for these results and so Mendel set about constructing a new theory of inheritance to explain this phenomenon. His theory posited an unobservable entity called a *gene* and it is these genes that cause the occurrence of characteristics in individuals. His theory of inheritance stated that:

- An individual has two genes for any one characteristic
- A parent will pass on one randomly selected gene onto its offspring
- If an individual has two different genes for the same characteristic, one of the genes will be *dominant* and one will be *recessive*. The individual will express the characteristics of the dominant gene.

The gene for purple flowers is dominant while the one for white flowers is recessive. Purple flowered peas have two identical genes coding for purple flowers whereas white flowered peas have two identical genes coding for white flowers. The second generation of peas therefore would inherit one purple and one white gene, since the purple flowering gene is dominant all second generation offspring have purple flowers. In the third generation each individual has a 25% chance of inheriting two purple genes, a 25% chance of two white genes and a 50% chance of having one purple gene and one white gene, resulting in a ratio of three purple flowered individuals to one white flowered individual in the third generation.

The state of gardening knowledge before Mendel's theory and the theory of evolution is offered by Botterill (1996 p109) as a fine example of a body of non theoretical empirical generalisations. Before these theoretical advances, gardener's knowledge consisted of a set of observations specific to

individual plant varieties and to general categories based on visible characteristics. When confronted with a new species our historical gardeners are not able to make predictions about its properties, its preferred conditions or the heritability of its characteristics. But would need to observe and record these variables for each new discovery in order to make future predictions about that species. As Botterill (1996 p109) notes “Like everyone else the gardener is entitled to form expectations that are hunches based on previous experience. But he is always a student of experience, rather than a theoretical pronouncer”.

Below I shall pick out several functional features of scientific theories and explain them with reference to Mendel’s theory of inheritance. I will then demonstrate how the body of information subserving folk psychological abilities shares these features and/or point out why theory would be favoured above a non-theoretical body of information by natural selection.

2.22 The positing of unobservables and implicit definition of terms

One of our core assumptions is that each observable phenomenon must in some way have been caused by some prior phenomenon. For any phenomenon therefore, there must have been a series of causal steps leading to its appearance. A theory is essentially a model of the causal system giving rise to the target phenomenon, that is what the theory is of, or about. A theory implicates a set of entities that are taken to be causally involved in the production of the target phenomenon and a set of rules⁷ of causal interaction between these entities. The causal interaction of theoretical representations therefore models the causal interaction of the entities that give rise to the target phenomenon.

Some of the entities that are implicated by the theory will be observable, they can be known independently of the theory and we therefore have a pre-existing concept of these entities which can be assimilated into the theory. The theory however may describe these entities in a novel way,

⁷ I use the term ‘rules’ in an abstract functional sense. It is not important to the argument how the rules of a theory are encoded. A theory can be realised in substrates as diverse as marks on a piece of paper, a microchip and – as I argue in this paper – a neural network (assuming physicalism)

laying down previously unknown causal interactions that it is possible for them to have. In this way a theory can build upon and enrich our concepts of observable entities. Not all phenomena however are caused solely by the interaction of observable entities, nor would we expect them to be. So in order for a theory to be an accurate model of the target system, it is often necessary to posit one or more unobservable entities and rules stating how these unobservable entities causally interact with each other and with the observable entities implicated in the theory.

With observable entities we saw it was possible to assimilate a pre-existing concept of that entity into the theory, perhaps enriching the inferential connections of that concept. Our knowledge of an unobservable posit however can consist of no more than its causal connections within that theory. Theoretical terms therefore gain meaning through *implicit functional definition* within a theory (Lewis, 1972). Our knowledge of a theoretical posit is as *something* that plays a certain causal role within a theory.

In the example above we can clearly see this structure. The phenomenon to be explained is the particular ratio of purple to white flowers in the second and third generation individuals in the cross breeding experiment. In order to describe the causal system Mendel found it necessary to posit an unobservable entity, the gene. The theory postulates that each individual plant possess two of these genes, that one of these genes is randomly selected to be passed to the plants offspring, and that they cause the observable characteristics of individual plants. We can know that they have certain properties, they can be either dominant or recessive. Our knowledge of the theoretical term gene therefore consists of nothing further than our knowledge of its causal role within the inheritance theory. What we don't know is anything about its intrinsic nature or structure, for genes this knowledge had to wait until the discovery of DNA which provided a suitable candidate for the functional role set out in Mendel's theory of inheritance.

Does our body of information posit implicitly defined unobservables? Yes, in order to accurately predict and explain others we make reference to unobservable mental states such as *belief, desire*

and *intention*. Our knowledge of these states consists solely of their interactions with environmental input, with each other and with output. It is important to note however that in order for mental states to gain meaning in this way TOM has to be consciously known. If the causal interactions of intentional states are known unconsciously as has been suggested by some theorists, then the meaning of mental states would also be unconscious. This is clearly not the case, we are consciously aware of the meaning of mental state terms. For this reason it is necessary to demonstrate in section 2.3 that TOM is a consciously known theory.

2.23 Prediction by projection

Theories enable predictions of a target system. For example the theory of inheritance could be used to predict the ratio of purple to white flowers in a repeat of the same experiment, but this could also be achieved using a simple empirical generalisation, we could merely note that when purple and white flowers are cross bred the resultant generations have a definite ratio of purple to white flowers and base a prediction on this. What marks a true theory out from generalisations of this kind is that the theory can be *projected* to make predictions of phenomena beyond those already observed. Using the theory of inheritance we could go on to predict the ratio of purple to white flowers beyond the third generation. We could also extend the theory to make predictions regarding inheritance of all visible characteristics controlled by a single gene. For an empirical generalisation to make such predictions it would need to be augmented with fresh observation. Empirical bodies of information are “open to extension, *but by addition not projection.*” (Botterill 1996 p109)

Does the body of information subserving folk psychological abilities allow prediction by projection?

Yes, on a daily basis we are able to make reasonably accurate predictions of the output of other humans, even those who we have not previously met. We regularly make predictions of output in novel situations in a way in which a body of empirical information would not allow. Take for example a situation that involved me living amongst a primitive people whose customs I knew nothing about. If I was to discover that the witch doctor *believed* that standing on his head would please his god,

and that he had the *desire* to please his god I would conclude that the witch doctor would regularly stand on his head, despite my ignorance of their religious customs.

2.24 Deep causal explanation

Theories can provide a deep causal story to *explain* the appearance or existence of certain phenomena. Using a specific empirical generalisation, the superficial explanation of the ratio of purple to white flowers would be simply because this ratio has happened to occur in the past.

Theory allows a deeper explanation by giving an account of the entities and their interactions within a causal system that result in the occurrence or existence of the target phenomenon. Indeed an explanation could be defined as giving an “abstract, coherent and causal account” (Gopnik & Meltzoff, 1997) of something.

Folk psychology allows the deeper type of explanation characteristic of theories. Taking again the example of the primitive people, I am able to explain the head standing with recourse to the causal interaction of witch doctor’s mental states. Reliance on a non-theoretical body information would only allow the superficial explanation that witch doctors tend to stand on their head if I had previously observed the ceremony on many occasions. If I was observing this tribal head standing for the first time I would be able to provide no explanation at all.

2.25 Cognitive Economy

A further feature that marks theories out from other bodies of information is their relative cognitive economy. As Botterill (1996) notes, “*Theories produce cognitive economy through integration of information in a small number of general principles*”. Unlike the three features discussed above, it is not possible to directly argue that the body of information subserving folk psychology possesses this feature. What we can do however is argue that the relative cognitive economy of theories provides a selection advantage to organisms and so theoretical bodies of information are likely to have been favoured by natural selection.

As noted above, theories can be thought of as an abstract model of the causal systems that give rise to a target phenomenon. Both the observable and the unobservable causally interacting entities of the target system are replicated by abstract causal representations of the theory. Cognitive economy is achieved because the same abstract causal structure can be applicable to a wide range of phenomena.

“The power of knowledge systems high in theoreticity to reduce the number of laws and principles needed to account for the data, replacing a large class of narrow-scope principles with a smaller class of more general ones” (Collin, 1985 p61)

Thus by recognising phenomena as tokens of a theoretical type, similar levels of function in terms of explanation and prediction to those gained by bodies of empirical information can be achieved using far less processing power.

For this reason it is likely that natural selection would prefer a theoretical body of information to a non theoretical one. The reason for this is that humans can accurately be described as cognitive misers (Stanovich 2009), that is, humans will use the minimum possible cognitive resources to achieve a task. Humans and their brains have been shaped by evolution and the advantages to the organism of increased processing power have to be balanced against the costs of any increase. Carruthers (2006) notes that chief among these costs are energy consumption and increased head size. The brain consumes approximately 20% of the body's energy, around 8 times the average energy consumption for an organ of its size. Any survival advantage conferred by increased processing power would be tempered by the higher levels of food that would need to be consumed in order to maintain such levels of processing. As noted above, similar levels of predictive and explanatory performance can be achieved using a theory with a relatively low processing demand to those achievable by a more fine grained and specific body of empirical information with a relatively high processing demand. In addition to the energy costs, increased processing power and thus a larger brain requires a larger head. This increases the probability of maternal mortality and in order

to minimize this risk, infants are born at an earlier stage of development therefore requiring a longer period of parental care before being able to fend for themselves. Theory therefore confers the same advantages at a lower cost to survival and so is likely to be the strategy settled upon by evolution.

2.26 Interpretation

The final feature that I will highlight is the interpretive function of theories. Theories are used to interpret evidence. They achieve this by providing an ontology, that is to say, they state which entities exist (both the observable and the unobservable entities of the theory). In addition to an ontology, a theory also provides a criteria which determines what counts as evidence for considering an entity that is perceived (or quasi perceived) to be identified with a representation of the theory. The same piece of evidence can count for many different conclusions, depending on the theory being deployed. This has been demonstrated by Kuhn (1996 [1962]) who analysed the history of science showing how the same evidence can be used to reach different conclusions depending on the paradigm (the framework of assumptions that underpins and directs scientific activity) that scientists were working in and also how some evidence can be ignored if it does not fit into the paradigm in which a scientist is working. As Gopnik and Meltzoff (1997 p37) note “Theory driven interpretations help solve what computer scientists call the frame problem. Theories provide a way of deciding which evidence is relevant to a particular problem”. In order to achieve this, *a theory must contain rules that enable the inference that we are perceiving a particular theoretical entity from an observed pattern of information.*

Theories provide a set of mental state types into which tokens can be subsumed in order to make sense of the world. In the words of Thomas Kuhn (1962 p24) the theorist attempts to “force nature into the preformed and relatively inflexible box that the paradigm supplies”. In other words, in addition to encoding the rules of interaction between theoretical entities, a theory determines the evidence that counts in favour of the subsumption of a particular token entity into a theoretically defined type entity.

Cognitive economy is achieved because we only need to know the rules that govern the causal interaction of a relatively small number of theoretically defined types rather than rules that govern the interaction of an impossibly large number of token entities that exist in the real world.

Possession of a TOM allows a human to economically find evidence that they themselves or a third person are currently in a particular mental state as defined by the theory. For this reason this interpretive feature of theories would provide the same benefits to the human organism as cognitive economy in general. It is my contention in this essay that the interpretive feature of theory plays a central role in the ascription of intentional states to oneself. This will be investigated in chapter 3.

I have demonstrated above that the body of information subserving folk psychological abilities displays all the hallmarks of a theory. Further to this that we have predictive/explanatory abilities that could not be explained if that body of information were structured in a sub theoretic way.

Finally that theoretically structured bodies of information provide benefits to the organism over and above those that would be provided by a non theoretical body of information and so natural selection would tend to favour a body of information structured theoretically. We are now able to conclude from this section's explorations that the body of information subserving folk psychology can be accurately described as a *theory of mind* (TOM). In other words our understanding of how the human mind works (both the minds of others and our own mind) is underpinned by a theoretical body of information which, like all theories, is composed of (i) a set of representations of the causal entities of the target system (both observable and unobservable), (ii) a set of rules governing their interaction and – most importantly for my overall argument – (iii) a set of rules determining the evidence necessary to infer the presence/occurrence of an entity of the target system. These theoretical components and their relation to ascription will be discussed further in section **2.5**.

2.3 The unenunciability objection and the erroneous appeal to tacit knowledge

In this section I raise the unenunciability objection to TOM theory and demonstrate that the traditional TOM theorists appeal to a tacitly known theory is inadequate. I achieve this by clarifying the tacit information/conceptual knowledge distinction and demonstrating how the two components that need to be construed as tacit information to avoid the objection are in fact part of our conceptual knowledge.

The objection can be summarised as follows: If our knowledge of folk psychology is subserved by a theory why are we unable to state its *laws*? As Goldman (1989, p167) states “Actual illustrations of such laws are sparse in number; and when examples are adduced, they commonly suffer from one of two defects: vagueness and inaccuracy.....But why, one wonders, should it be so difficult to articulate laws if we appeal to them all the time in our interpretive practice?”

A popular response to this objection is to appeal to a tacitly known theory of mind. We are unable to state the laws of the theory because they are not consciously known to us. I will argue in this chapter that this manoeuvre is not a satisfactory response to the objection. At the end of section 2.2 we concluded that TOM contains the following components: (i.) a set of representations of the causal entities of the target system (both observable and unobservable), the causal interaction of which are taken to give rise to the target phenomenon, (ii.) a set of rules governing the interaction of these entities, giving meaning to them (either by enrichment of current concepts or by the definition of new ones⁸) and (iii.) a set of rules determining the evidence necessary to infer the presence/occurrence of an entity of the target system. Clearly if (iii.) were tacit while (i.) and (ii.) are consciously accessible we would not avoid the objection as it is knowledge of (ii.) that Goldman is demanding. But if we have conceptual knowledge of (ii.) then we must also have conceptual

⁸ See section 2.22

knowledge of (i.) because the entities of a theory are implicitly defined by their interactions.

Therefore (i.) and (ii.) must stand or fall as conceptual knowledge together.

In section **2.31** I shall draw a distinction between tacit information and conceptual knowledge. In

section **2.32** I will show why we have conceptual knowledge of TOM components (i.) and (ii.)

2.31 Tacit information and conceptual knowledge

Many organisms can respond differentially to a variety of stimuli, and so must contain a body of information matching stimuli to response. I contend that such bodies of information can be divided into two categories: tacit information and conceptual knowledge. If we were to describe all bodies of information subserving differential responses as *knowledge*, many types of behaviour that we would intuitively consider as not exploiting knowledge would be classified as such. For example types of differential response such as the process of phototaxis whereby a maggot moves away from a light source could be interpreted as exploiting a body of knowledge. Phototaxis causes a maggot to burrow away from a light source thus reducing the probability of it being noticed by a predator. But the maggot has no *knowledge* of this. Such a process could even be said to involve simple concepts⁹. What then, marks out true conceptual knowledge from a mere body of information such as that possessed by the maggot?

Consider an example of a differential response: the verbal report “I see a red ball”¹⁰. Current technology allows the creation of robots that can be trained to respond differentially with a verbal report to a variety of stimuli. When the robot asserts “I see a red ball” when presented with a red ball, it could be said to be relying on a body of information regarding what objects are called. But it does **not** have the *conceptual knowledge* that what it is observing is a red ball. The difference between this type of verbal report and the type made by a human is that the human knows the

⁹ Carruthers’ (2006) describes how bees exploit basic concepts in order to navigate to and from the hive.

¹⁰ What Sellars (2007 [1954]) would call a “language entry transition”

inferential role of the concept “red ball” in what Sellars (1956 p76) called the “space of reasons”. The human is able to take the concept “red ball” and - in conjunction other conceptual knowledge - make *conscious* inferences to support other propositions not directly related to the original differential response, in this way conceptual knowledge is said to be *inferentially integrated* and *consciously accessible*. Conversely, bodies of information and concepts that are unable to support such conscious inferences are said to be *inferentially encapsulated* and *inaccessible to consciousness* (Stich 1978) such bodies of information will be referred to as tacit.

The distinction can be seen in a linguist’s knowledge of transformational grammar. While a linguist has conceptual knowledge of transformational grammar, their actual ability to produce and discriminate grammatical sentences is subserved not by this knowledge but by a body of tacit information “A linguist might possess all the concepts involved in transformational grammar, but this does not suffice to show that their knowledge of transformational grammar is conceptual. Indeed given that linguists acquire their ability to produce and understand grammatical sentences separately from their linguistic concepts, there are good reasons to suppose that it is not.” (Maibom 2003). It can be said that the linguist has two separate bodies of information regarding transformational grammar, one of the tacit variety that was gained while learning to use a language and a second of the conceptual variety learned during his/her academic studies of linguistics. I shall reserve the term knowledge for bodies of information of the conscious conceptual variety, in this way we can avoid the consequence of attributing maggots or differentially responsive robots with knowledge.

If an individual possesses knowledge of a concept i.e. if that concept is consciously accessible and inferentially integrated, then that individual is able to understand that concept. If an individual has a tacit concept they do not understand its content (Maibom 2003). In order to understand the content of a proposition, that proposition’s content must be part of our conceptual knowledge and not part of a body of tacit information.

To illustrate this consider the linguist discussed above and another individual who has had no academic linguistic education. Both individuals are able to accurately produce and judge grammatical sentences of their native language, but only the trained linguist would understand a description in terms of linguistic concepts of how such judgements are achieved. The fact that the layperson is equally as competent at grammatical judgements as the linguist, despite his lack of conceptual knowledge of transformational grammar, further demonstrates that grammatical judgement capacities are subserved by a body of tacit information and not a body of conceptual knowledge. As Maibom (2003) notes “Linguistic research illustrates the inferential encapsulation of knowledge [that is information not true knowledge in my terms] of grammar well. Linguists are competent speakers, but they cannot use their grammatical knowledge as such in their research. They are compelled to observe and theorise.” In summary, the fact that lay people are able to produce grammatical statements and yet do not understand linguistic concepts demonstrates that grammatical judgements are subserved by a body of tacit information rather than the conceptual knowledge possessed only by the trained linguist.

To conclude my clarification of this distinction, tacit information is inferentially encapsulated and inaccessible to consciousness. Even though it may incorporate tacit concepts, we do not understand the content of these concepts when it is presented to us. Conceptual knowledge on the other hand is inferentially integrated and accessible to consciousness. When presented with the content of a concept of which we have knowledge, we are able to understand that content.

2.32 Conceptual knowledge of TOM

We are now in a position to determine whether the first two components of TOM are tacit information or conscious knowledge. If those components can be shown to be inferentially integrated and consciously accessible then it would seem appropriate to assign them to the category of conceptual knowledge as opposed to a body of tacit information as they have been traditionally construed by TOM theorists.

The following examples of generalisations are given by Churchland (1981) as examples of tacit folk psychological laws:

Generalisation A - $(x)(p) [(x \text{ fears that } p) \rightarrow (x \text{ desires that } \neg p)]$

Generalisation B - $(x)(p) [(x \text{ hopes that } p) \& (x \text{ discovers that } p) \rightarrow (x \text{ is pleased that } p)]$

Generalisation C - $(x)(p)(q) [((x \text{ believes that } p) \& (x \text{ believes that (if } p \text{ then } q))) \rightarrow (\text{barring confusion, distraction etc., } x \text{ believes that } q)]$

On reflection however, these laws do not seem to be tacit since when put into ordinary language most adults would understand them, although not necessarily agree with them. This fact seems to suggest that these statements are inferentially integrated, that is we are able to understand and assent to their content, and so are candidates for being part of our conceptual knowledge.

Another fact that counts against the claim that our knowledge of folk psychology is tacit is that we are able to integrate learned knowledge into the folk psychological framework and use this new knowledge in the prediction and explanation of behaviour, thus supplementing and enriching our folk psychological capacities (Maibom 2003). Take for example knowledge of the Langer effect discussed above, the predicting group were obviously unaware of the little known Langer effect, but it is highly likely that an individual who was taught details of the effect would have provided a more accurate prediction than those who were ignorant of it. The taught individual would have assimilated the information of the effect into his conceptual knowledge and used it to improve the accuracy of his behavioural prediction. This is just one example of taught psychological knowledge that can be integrated into and enrich core folk psychological information and it is possible to think of many more. Learned information such as the Langer effect is a type of conceptual knowledge and the fact that this knowledge can be integrated into folk psychology shows that theoretical components (i & ii) are inferentially integrated and different in kind to the type of tacit information that subserves grammatical judgements.

Above it has been shown that at least components (i.) and (ii.) of TOM are inferentially integrated conceptual knowledge and not tacit information. We would therefore expect that TOM would be consciously accessible. But the original motive for classing theory of mind as a tacit theory was that we seem unable to recount the laws and generalisations of the theory (although we do understand when presented with their content). Does the fact that we are unable to recount many folk psychological laws pose a problem for the idea that folk psychology is a theory? I suggest that it doesn't because our folk psychological knowledge is not of laws and generalisations as suggested by theorists such as Churchland (1981) and Botterill (1996) but of theoretical models and hypotheses as suggested by Maibom (2003) and Godfrey-Smith (2005). This response to the unenunciability objection will be considered in the next section.

2.4 Theories as models and hypotheses – A response to the unenunciability objection

In this section I will sketch what I propose is a satisfactory response to the objection levelled the TOM account by Goldman (see section 2.3)

The thrust of my argument is as follows: Any generalisation that I verbalise is likely to be, in general, false as it can only be applied in certain situations as determined by the hypotheses that accompany the model. The reason I am unable to recite the laws of TOM is therefore because TOM contains no such universally applicable laws. If TOM is structured in this way, we would not expect people to be able to recite its laws and the objection misses the point. In Section 2.41 I will describe the models and hypotheses account of how scientific theories are structured (Giere 1988, 1999). In section 2.42 I will detail the consequences of this view for TOM and the unenunciability objection.

2.41 Scientific theories as models and hypotheses

Traditionally folk psychology has been construed as a collection of generalisations and laws stipulating the causal relationships between intentional states (see for example the laws suggested by Churchland (1981) above). In this section I will show that the actual structure of theories is more

accurately described as collections of models and hypotheses, in section 2.5 I will show how this structure fits with my list of theoretical components.

The motivating factor behind describing theories as models and hypotheses is an inconsistency in their description as consisting of general laws. The problem with the construal of theory as a body of laws is that as generality increases the further the law moves away from an accurate representation of reality, as Maibom (2003) notes “Either laws of science are general and not true (because at best approximately true), or very specific and true”. This inconsistency can be rectified by re-describing scientific theories as models and hypotheses.

What does it mean to say that theories consist of models and hypotheses? The view can be explained taking Giere’s (1988) example of the harmonic motion of a simple pendulum. The equation $m \frac{d^2 x}{dt^2} = - (mg/l) x$ is used to understand the motion of the pendulum, but the point is that this is not true any real world pendulum but only true of an idealised model of a pendulum. The forces acting upon real world pendulums are far more complex, for example the model does not take into account other factors such as friction or the variability of gravity across the swing. The equation, while true of the idealised model of a pendulum is strictly false of any real world pendulum. As Maibom (2003) notes “It is therefore useful to think of science not as providing bodies of information consisting of laws that describe states of the world, but as providing scientists with models by means of which they can understand the world”. While never perfectly true of any real world system, a theoretical model varies in degrees of appropriateness of application to real world systems. Theoretical hypotheses concern the relation of models and real world systems and state the respects in which the model matches or fits a real world system. So while hypotheses can be described as true or false it is not appropriate to describe models as true or false. Giere (1988) uses the example of a scientist’s explanation of the motion of the earth and the moon, the theoretical model in question is a two particle Newtonian model with an inverse square central force. A

hypothesis states the degree of fit between the real world system (the earth and moon) and the theoretical model (the two particle Newtonian system) and so would be (Giere 1988 p81):

The positions and velocities of the earth and moon in the earth-moon system are very close to those of a two-particle Newtonian model with an inverse square central force

The above hypothesis states the properties that the real world system shares with the theoretical model, namely the positions and velocities of the earth and moon. Yet the earth-moon system has many properties that are not included by the scientific model such as the chemical composition of the two bodies and the fact that the system is part of a yet larger system of moving bodies. The hypothesis also states the degree that which the specified properties of the real world system fit the properties of the model, in this case it is stated that they are 'very close' meaning that the properties of the real world system not taken into account by the model – such as the gravity of other bodies within the larger system - would have negligible effects on any predictions made of the real world system using the idealised model. In order to gain the functional advantages of theories listed above a scientist requires both an idealised model and a hypothesis stating the degree of fit between the model and the target real world system. The proposed structure can also be seen in the example of the theory of the Mendelian theory of inheritance described above, the idea of the interaction of individual genes is an idealised and much simplified model of the actual process of inheritance. Genes are not discrete packets of information as suggested by Mendel's theory. The actual medium of inheritance, as demonstrated in the mid part of the twentieth century, are long strings of DNA molecules. Each gene actually consists of many different pieces of DNA that are spread, seemingly at random over the entire length of a string of DNA. Genes may share sections of DNA, some pieces of DNA may be inactive in the absence of other certain pieces of DNA, as the geneticist Richard Dawkins (1976 p24) explains:

The manufacture of a body is a cooperative venture of such intricacy that it is almost impossible to disentangle the contribution of one gene from another. A given gene will have many different effects

on quite different parts of the body. A given part of the body will be influenced by many genes, and the effect of any one gene depends on interaction with many others. Some genes act as master genes controlling the operation of a cluster of other genes. In terms of the analogy, any given page of the plans makes reference to many different parts of the building; and each page makes sense only in terms of cross-references to numerous other pages (Dawkins 1976 p24)

So in order to make sense of inheritance, without analysing the incredibly complex interaction of individual DNA base pairs, we can use the theoretical model of Mendelian inheritance to make predictions, explanations and interpret evidence in a highly cognitively efficient way. The accuracy of this model is comparable to that which could be achieved by a far more fine grained analysis in terms of base pair interactions, without the vast processing power the latter would demand.

It might be thought that theoretical models are limited to the advanced and specialist type of knowledge that is science. But a closer look reveals that any *type* analysis involves simplified models of target phenomena. To make any prediction of a token of a particular phenomenon it is necessary to assimilate that token to a particular type, whether it be a scientific prediction of planetary motion or more everyday predictions, such as the prediction that upon entering a room I have never previously entered, pressing a switch on the wall will cause a bulb will illuminate. The second example may seem trivial, but I see the switch on the wall as a token of a particular type, that of light switches and I know that the pressing of a light switch generally results in the illumination of a bulb. Switches and bulbs can vary in design greatly and I need not know the exact details of the system encountered, but I am quickly able to assimilate any token of a switch into my simple model of a light switch/ bulb system and make a prediction on that basis. Models vary in complexity, from the highly complex mathematical models of theoretical physics to simple models of systems that a modern human will encounter on a daily basis. The human ability for prediction, explanation and interpretation based on the assimilation of a novel token into a previously known type involves idealised models of some kind.

2.42 TOM as models and hypotheses

Take generalisation A above (section 2.32), essentially this states that if an agent X fears that P, X desires that not P. It is easy to think of counterexamples to this generalisation, take someone with severe toothache who fears trips to the dentist. Despite their fear of the dentist, their overriding desire is to end the pain of the toothache and so contrary to A, in this particular situation: $(x)(p) [(x \text{ fears that } p) \& (x \text{ desires that } p)]$. It is easy to multiply such examples; it is not uncommon for someone to fear something but to have an overriding desire that causes them to desire that thing. If this is a law of human behaviour then it is not a very good one, as in many situations it does not apply. Counter examples can also be provided to the other laws in Churchland's list. The generalisations in Churchland's examples can now be seen to be highly simplified models of real world agents. They represent the agent as having only a handful of mental states that interact resulting in output. However any real world agent has dozens, perhaps hundreds of mental states, all of which are potential factors in the resulting output. Compare this with Giere's (1988) example of a two particle Newtonian model that abstracts away from factors that are assumed to have a negligible effect on the interaction and the similarity is clear. Churchland's examples of generalisations are true of the entities defined in idealised models but not of agents in general. Despite not being true of any real world agent, mental model generalisations are evidently very effective, as our folk psychological capacities are on the whole very effective. The reason behind this is that, in addition to knowledge of how theoretical model entities interact with each other, we have additional knowledge in the form of hypotheses dictating the correct application of each model. In the case of folk psychology hypotheses, we would need to determine how applicable each model is likely to a particular real world agent. It has been proposed by Maibom (2003) that folk psychological hypotheses are formed by background knowledge of agents with whom the attributor has previously interacted and of stereotypes in the case of agents with which we have not previously interacted and in the cases where none of this information is available the most probable model would be selected. Consequently, the theoretical model account would predict that the more background

knowledge available to the attributor about an agent, the more accurate attributions and hence predictions are likely to be with respect to that agent. This can be clarified with an example provided by Maibom (2003). Consider the following: If and agent P desires that Y and believes that performing X will achieve Y, then P will perform X. This is not as been previously thought a law of TOM, but a theoretical model. Suppose that the Y that P desires is 'money' and that the X that P believes will achieve this is 'robbing a bank'. We can see that in this instance the model:

$(P)(X)(Y) [((P \text{ desires } Y) \ \& \ (P \text{ believes that } (if \ X \ \text{then } Y)) \rightarrow (P \text{ will perform } X)]$

Would (hopefully) not be true of the vast majority of the population, although may be true of a small proportion of it. This is due to the interference of factors not accounted for in the model, namely the threat of punishment and/or moral objections.

How then are we to know when this model is applicable to P? If we had not previously met P then it is likely that our default position would be that the model did not apply, because - on the balance of probabilities - it is more likely that a person would not rob a bank to achieve wealth. What if we knew the individual is from a social group in society that is stereotyped¹¹ as having criminal dispositions? Perhaps if we were a particularly prejudiced individual would then we would feel it was appropriate to apply the model in question. What if we knew that P had a history of crime and that he had been punished on many occasions previously? In this case all but the most unprejudiced individuals would feel that the model was applicable to P. This example makes clear that background knowledge (consciously or unconsciously) affects the decision as to whether a theoretical model is applicable to a particular person in a particular situation. What the example and the preceding discussion also makes clear, is that there are no such things as 'laws' when it comes to TOM. If the overarching argument of this essay is correct then we would make decisions of the applicability of models not only to others, but to ourselves. Our background knowledge of ourselves is far greater

¹¹ Such stereotypes - as with the 'laws' of TOM - are generally false.

than our background information of any other person, making swift, unconscious model applicability judgements to ourselves both natural and extremely accurate.

So, we are finally in a position to successfully counter the unenunciability objection. Why can we not state the laws of TOM? It is because there is no such thing. TOM contains no universally applicable models and so when asked to state such laws, we draw a blank.

2.5 Components of theory and ascription

In this section I add a fourth component to finalise the list of theoretical components. Namely theoretical hypotheses discussed in the above chapter. I discuss each of the components roles in forming a theoretical prediction/explanation. I identify the third interpretive component as the mechanism by which we ascribe intentional states to others and ourselves. The final list of theoretical components is listed below:

(i.) A set of entities (both observable and unobservable), the causal interaction of which are taken to give rise to the target phenomenon,

(ii.) A set of rules governing the interaction of these entities, giving meaning to them (either by enrichment of current concepts or by the definition of new ones)

(iii.) A separate set of inferential rules determining which entities of perception are to be identified with the entities of the theory or which phenomena of perception are indicative of the presence of a theoretical entity.

(iv.) A set of hypotheses that determine the applicability of a theoretical model in a particular situation.

Theoretical explanation/prediction therefore proceeds in the following way. Firstly, the presence or occurrence of an entity that is part of the theories ontology is inferred from the available evidence using the rules laid out by (iii.). This provides the theory with an input. Secondly the most likely

model (consisting of (i.) & (ii.)) is selected from a bank of models consistent with the input using the rules determined by (iv.). Finally, the appropriate model determines the most likely interaction of the entities implicated in the first step. In the case of prediction the model provides a means of inferring the likely output. For an explanation the model provides a description of the causal steps leading to the occurrence of a particular output.

This section has established that folk psychological abilities are subserved by a TOM. Folk psychological predictions/explanation therefore proceeds in the way described above. The initial ascription of intentional states must therefore be achieved by an interpretive mechanism equivalent to theoretical component (iii.). I shall henceforth refer to component (iii.) of TOM as the TOM ascription mechanism.

This section brings to a close chapter 2 of the essay. And so it is an appropriate time to summarise what we have concluded. In section 2.1 we concluded that folk psychological abilities are driven by information rich processes, that is, they are subserved by a body of information which determines the rules of interaction between intentional states. In section 2.2 we concluded that this body of information subserving folk psychology is structured theoretically and can therefore accurately be described as a TOM. Like all theories TOM must consist of at least three separate components: (i.), (ii.) and (iii.). In section 2.3 we introduced the unenunciability objection and showed how the appeal to a tacit TOM is an unsatisfactory response to the objection. In section 2.4 we showed how TOM does not contain the laws that the objector assumes and for this reason the objection loses its bite. We also added a further theoretical component (iv.). In section 2.5 I discussed the role of each of these components in forming an explanation/prediction. The overall conclusion of chapter 2 can now be stated. Folk psychology is subserved by a theory of mind. Theories, of which TOM is an example, consist of four basic components: (i.), (ii.), (iii.) and (iv.) as defined above. Of which, at least (i.) and (ii.) are accessible to consciousness. Initial ascriptions are achieved by means of component (iii.) which determines how evidence is interpreted in order to infer the presence or occurrence of an

entity that is part of the theories ontology. Initial ascriptions of intentional states are therefore achieved by component (iii.) of TOM, that is, the TOM ascription mechanism.

In the next chapter I will examine component (iii.) of theories. I will show that theoretical interpretation can proceed without any conscious knowledge of inference therefore demonstrating that theoretical self-interpretation is consistent with the *phenomenological directness* of self-ascription. I will also describe the information available to the TOM ascription mechanism that enables self-ascription even in the absence of overt behaviour.

3

THEORETICAL INTERPRETATION

With the theoretical nature of intentional states established, in this chapter I shall describe in section 3.1 how theoretical self-interpretation is consistent with the seemingly direct nature of our knowledge of intentional states by describing examples of theoretical interpretation in which we have no conscious awareness of inferring. I will firstly draw a distinction between *observation* and *perception* and secondly describe several examples in which theoretical posits can be directly perceived through a process of *theory-laden perception*. In section 3.2 I will show what information the TOM ascription mechanism has access to and how this enables the theoretical self-interpretation account to explain instances of self-ascription in the absence of overt behaviour.

3.1 Theory-laden perception

Ascribing intentional states to oneself is an everyday activity. Each day, we acquire the knowledge that we *believe* that X, *intend* that Y and *desire* that Z. When we do self-ascribe, it is a fast, unconscious process. We are certainly not conscious of any *theorising* as we make no conscious inferential steps. This might be thought to be an objection to TOM. Can a self-interpretive theory account for this *phenomenological directness* of intentional state self-ascription? In this section I intend to demonstrate that it can, by showing how the possession of a theory can allow the direct phenomenological perception of unobservable theoretical posits. I suggest that this kind of *theory-laden* direct perception is the process by which we acquire non-inferential¹² knowledge of our intentional states.

Firstly, I will clarify the usage of the terms *observe* and *perceive* which I will be using in the following discussion. Consider one of the head standing tribesmen discussed earlier. He has lived deep in the

¹² That is by no conscious theoretical inference

Amazon for the whole of his life. Suppose – to show gratitude for letting me stay with him - I invite this tribesman to come and stay with me for a couple of weeks at my home in England. During his stay we take a trip to the zoo. As we approach the hippopotamus enclosure our gazes fall upon the beast within. While I am quite familiar with hippopotami - having previously seen them in the zoo and having watched several BBC documentaries featuring them - my new tribesman friend has never seen one before (hippopotami being indigenous to Africa). While we both *observe* the hippopotamus, only I *perceive* a hippopotamus. *Perception* unlike *observation* involves the acquisition of knowledge (see section 2.3 for a definition of knowledge). In this case I perceive a hippopotamus, that is to say, I observe a hippopotamus and acquire the attendant knowledge that what I am observing is a token of the type *hippopotamus*. After listening to my brief explanation and reading the information board the tribesman acquires the concept of a hippopotamus. On a second visit to the zoo, we once again approach the hippopotamus enclosure and – to our mutual satisfaction - we both at once *perceive* that it contains a hippopotamus. The point is that while we can observe something without having a pre-existing concept of that thing we cannot perceive it. We cannot perceive X unless we have a concept of X.

I will now describe an example of what I contend is the theory-laden perception of theoretically defined states. Consider the following example from Gopnik (1993) “Master chess-players report that they no longer see the board in terms of individual pieces and squares but as a set of competing forces and powers. They need not calculate that an isolated king is vulnerable; they see he is.” In other words, the chess-master directly perceives the strength of the king. There are two processes by which this direct perception can be achieved, theory-laden perception and direct causal perception. I will argue that the chess-master’s perception is achieved by the former of these two processes. And so while this knowledge is interpretive in nature it can be gained directly, without the need for conscious inference.

Before embarking on this task however, it is necessary to clarify the two types of perception. Firstly theory-laden perception, which - despite its phenomenological directness - is interpretive in character and dependent on the possession of a body of information regarding the object perceived. Theory laden-perception allows the *perception* of unobservable theoretical entities which – by their very nature - cannot be directly *observed*. Their existence therefore needs to be inferred through the observation of indirect evidence. Theory-laden perception is therefore a species of theoretical interpretation in which theoretical component (iii.) is tacitly known and the interpretive inferences are made unconsciously. It is this fact that accounts for the phenomenological directness of the theory-laden perception of an unobservable entity.

The second type of perception is direct causal perception. This type of perception is achieved by a relatively direct channel in that it involves no interpretive inferences, conscious or otherwise. Because of the lack of inferential steps, this method is relatively reliable when contrasted with theory-laden perception whose reliability correlates with the accuracy of the body of information on which it relies.

I will argue that the chess-masters perception of the strength of the king is achieved through theory-laden perception. Firstly by showing that knowledge of the kings strength is achieved by the beginner through a series of inferential steps, suggesting that the master makes the same steps although unconsciously. And secondly, by describing a thought experiment which suggests that the chess-masters perception is unlikely to be of the direct causal variety, as it can be shown to be unreliable.

In contrast to the chess-master, in order to gauge the strength of the king a chess beginner needs to consider hypothetical moves, to run through several scenarios in order to consciously calculate the level of threat to the king. In other words, the beginner gains knowledge of the king's strength by consciously inferring it from the observed evidence (the positions of the pieces on the board). It is unlikely that in making the transition from beginner to master the chess player changes from an

inferentially based knowledge to a non-inferential direct causal perception of the strength of the king. It is far more plausible that the beginner's inferential steps become so familiar that they can be achieved unconsciously. Indeed, while direct causal perception seems plausible for the perception of basic, observable objects¹³, it seems far less plausible for abstract, unobservable entities such as the strength of a king in a chess match.

That the chess masters knowledge is achieved through theory-laden perception is also supported by the following thought experiment. Suppose a genius of chess invented a new offensive strategy, one not previously considered by his fellow masters. The genius – with his new strategy a well kept secret - meets one of his fellow masters in a tournament. As the game progresses the opponent notes that the genius is employing an unfamiliar strategy, but observing his king he perceives that it is in a strong position and makes his moves accordingly. Later in the game as the strategy unfolds, the genius springs his trap. What the opponent perceived as a strong king is suddenly placed in an impossible position and within a few moves the genius mates his opponent. The new strategy is a success. The mode of perception with which chess masters perceive the strength of a king is clearly not a direct, reliable monitoring as in the second sense of perceive. For during the moves immediately prior to the springing of the trap, what the opponent perceived as a strong king was actually a vulnerable one.

The most plausible explanation for the chess-master's direct phenomenological perception is therefore a form of theory-laden perception. Through experience the chess master constructs a body of information¹⁴ about the game and it is this body of information that shapes his concept of 'a strong king'. The inferential steps which start from the observation of pieces on the board and which result in the perception of the strength of a king must therefore be mediated by master's body of chess information, for it is this which defines what a strong king is. The novel strategy employed by

¹³ Even this is disputed, see Sellars (1956) and more recently Rosenthal (2005)

¹⁴ This body of information may or may not be truly theoretical. Despite its name it is possible that theory-laden perception could be achieved by the exploitation of a sub-theoretic body of information. I will therefore leave open the question of whether the chess-masters body of chess information is truly theoretical.

the chess genius alters the states of play in which it is appropriate to assert “this king is in a strong position”. Thus, when the strategy becomes well known it would be accurate to say that chess masters’ perception is altered.

I have demonstrated above how possession of a body of chess information allows the phenomenologically direct theory-laden perception of the strength of a king. But it is not just our fortunes in chess that we can directly perceive in this way. Kuhn (1962) describes how a scientific education changes what an individual perceives “Looking at a contour map, the student sees lines on paper, the cartographer a picture of a terrain. Looking at a bubble chamber photograph, the student sees confused and broken lines, the physicist a record of familiar subnuclear events. Only after a number of such transformations of vision does the student become an inhabitant of the scientist’s world, seeing what the scientist sees and responding as the scientist does.” (Kuhn 1962 p111)¹⁵. As the science student learns a theory he acquires new concepts in the form of the implicitly defined posits of that theory, therefore increasing the number of entities which he is able to perceive. As with the chess example, that such perception is theory-laden as opposed to causal is demonstrated by the fact that such perception can be radically false. Take Kuhn’s (1962) classic example of the late 17th/early 18th century phlogistic theory of combustion. The theory posited an unobservable substance known as phlogiston. Combustion was held to be the process whereby phlogiston was removed from a substance and absorbed by the air. On observing a burning substance the 18th century chemist perceived the substance losing phlogiston. On making the same observation the modern chemist perceives the substance gaining oxygen. The 18th century chemist’s perception of phlogiston was obviously not directly *caused* by the presence of phlogiston, as modern chemistry tells us there is no such substance.

The above discussion demonstrates that direct phenomenological perception can be achieved through exploitation of a theory and/or a sub theoretic body of information. It shows that if no

¹⁵ This line of thinking is foreshadowed in the work of Sellars (1956) who argued that in order to perceive something – even something as basic as a red triangle - we must have a pre-existing concept of that thing.

conscious inferences are made in the acquisition of knowledge from observation, it does not mean that the process of acquiring said knowledge is non-inferential. It may be that those inferences are made unconsciously. The chess master's and the scientists direct phenomenological theory-laden perception challenges the intuition that such perception must be achieved by a form of direct casual perception.

The direction of this argument should now be clear. As demonstrated in chapter 2, our intentional state concepts are implicitly defined by our TOM. It is therefore possible to *perceive* our own intentional states through a process of theory-laden perception. The possession of a particular intentional state can unconsciously inferred from observation. Theory-laden perception therefore provides the self-interpretive theorist with a method of intentional state self-ascription that is interpretive rather introspective in character.

My usage of *observation* in this essay will not be limited to visual information as in the above examples. My definition of perception as the acquisition of knowledge means the observation of many different information channels can result perception, for example my tribesman friend could learn to perceive the hippopotamus by its feel or its smell. Moreover, perception need not be limited to objects or states of affairs in the external world. Indeed on the broad definition of perception above both introspective and self-interpretive theorists would agree that we *perceive* our own intentional states. The difference between the two positions lies in the type of information that is observed resulting in the perception of intentional states. The term observation is therefore also used in a broad sense, and so to say that we observe our overt behaviour is also to allow that both visual and proprioceptive information are available to the observing system.

This method of self-ascription has been proposed by Gopnik (1993). Gopnik suggested that self-ascription of intentional states was achieved by the observation of our own behaviour perhaps in addition to an undefined "Cartesian buzz" (Gopnik 1993 p11) of internal activity.

Gopnik's account however came in for heavy – and valid – criticism. Both Goldman (2006) and Nichols and Stich (2003) point out that the observation of behaviour is not a plausible alternative to introspection. The reason for this is that we have the ability to self-ascribe intentional states in the absence of observable behaviour. Sitting still at my desk deep in thought, I am able to ascribe to myself the *intention* to brew yet another cup of coffee. Yet I exhibit no coffee seeking behaviour - yet. In order for theory-laden perception to be a plausible alternative to introspection it must be able to explain self-ascription in situations such as this. This will be my task in the next section where I will describe what information – in addition to behavioural observation – is available to the self-interpretive mechanism.

3.2 What is observed when we perceive intentional states?

In this section I will describe the information available to the interpretive mechanism. While behavioural observation may be able to account for many cases of self ascription it cannot account for all of them¹⁶. If we are to accept the strong conclusion that all self-ascription is achieved by self interpretation then there must be other information available to the interpretive mechanism. For self-ascription to be described as self-interpretive the ascribing mechanism must perceive on the basis of observation of “the subjects current circumstances, or the subjects current or recent behaviour, as well as any other information about the subject’s current or recent mental life.” (Carruthers 2009).

I suggest – following Carruthers (2009) - that self-ascription could be achieved if the self-interpretive mechanism had access to *globally broadcast* information. The idea of a global broadcasting architecture has been developed and supported by the cognitive scientists Bernard Baars (e.g Baars *et al* 2003) and Stanislas Dehaene (e.g Dehaene & Naccache 2001). Essentially incoming sensory information is partially processed and then globally broadcast to systems geared to receive such information. These include conceptual, long to medium term memory, action planning and desire

¹⁶ Think again of my intention to brew coffee

forming systems (Carruthers 2006 p 84-85). I will argue below that knowledge of our own intentional states is based on this globally broadcast information. That is, the same information available to belief, desire and intention forming systems and **not** on the basis of the output of these systems themselves. I propose that this globally broadcast information fills the role of Gopnik's Cartesian buzz, enabling a self-interpretation account to explain self-ascription in the absence of overt behaviour.

So what information is globally broadcast and can access to this information explain how we can self-ascribe even in the absence of overt behaviour? In addition to partially conceptualised visual information (see Kosslyn 1994) and proprioceptive data giving information about our overt behaviour, globally broadcast information includes patterns of attention and visual and auditory imagery (Carruthers 2009). That is, "images" available to consciousness, for example sentences verbalised in inner speech; or visual images seen in what is commonly known as the "mind's eye". Since such imagistic information involves the same resources involved in actually hearing speech or actually seeing an image this information can be globally broadcast with other sensory information and hence would be available to the TOM ascription mechanism geared to receive such information (Carruthers 2009 p 124).

With access to such globally broadcast information we are now able to give an account of self-ascription in situations where we exhibit no overt behaviour. Take my self-ascription of the *intention* to brew a cup of coffee while sitting quietly at my desk. There are now many ways in which I could self-ascribe this intention; perhaps I had just verbalised the intention in inner speech or formed the mental image of myself downstairs boiling the kettle; perhaps I notice my attention repeatedly turning to my empty cup. Observation of any of these pieces of information individually or a combination of all of them could easily form the basis of my self-ascription of an intention. Such information is the equivalent to the pieces on the chess master's board or the lines on the physicists

bubble chamber photograph. *The ascription mechanism's observation of this globally broadcast information results in the theory-laden perception of a TOM defined intentional state.*

To some, giving the ascription mechanism access to information such as that described above might suggest that the ascription process can no longer be described as interpretive and rather counts as a form of introspection. For example, if we are able to verbally report our intentional states doesn't giving the ascription mechanism access to such reports beg the question? Don't we need *knowledge* of our intentional states in order to report them? I suggest that we don't. I do not dispute that one of the fastest ways to find out if you believe X is to ask yourself "do I believe that X?" This method of self ascription –known as self-assent - has been suggested by theorists as diverse as Gordon (1996) and Carruthers (1996). What I do dispute is the claim that self ascription must occur in order for a verbal report to be made. As noted above (Section 2.3) a verbal report is a species of differential response. And differential responses are caused by unconscious tacit information rather than conscious conceptual knowledge. The robot described above does not need to *perceive* the ball in order to report its presence. Similarly humans do not need to *perceive* an intentional state in order to report it, *observation* will do. Indeed, what I am suggesting is that self-ascription can be made on the basis of an unconsciously made verbal differential response. In other words, in cases of self-assent we gain knowledge of our intentional states by hearing ourselves report them. Such verbal reports need to be *observed* and *interpreted* by the TOM mechanism in order for self-ascription to occur. Self-assent can therefore be plausibly construed as a process of self-interpretation in which a differential response is first elicited and then interpreted. The objector may press further however. Doesn't giving the TOM mechanism access to verbal reports represent a direct and reliable link, the type of link characteristic of introspection, not interpretation? However, the globally broadcast imaged sentence would need to be *interpreted* in the same way as the utterance of a third person. This requires information such as context and tone to be taken into account in deciding on the correct interpretation of the sentence. The applicability of a particular TOM model to a subject (see section 2.4) also partially informs ascription (is she the type of person to believe this?) and so self-

image and memory are also likely to inform self-ascription (am I the type of person to desire that? consider the intention to get a coke in section 1.1). This integration of various pieces of information is what is distinctive about theory-laden perception. The chess masters knowledge based on the integration of information of the positions of many pieces on the board. The physicist the many swirling patterns of bubble chamber photograph. The TOM ascription mechanism integrates the various pieces of globally broadcast information with memories and other stored information and produces the best self-ascription on the balance of evidence. As with other examples of theory-laden perception, theoretical self-interpretation on the basis of such diverse sources of information admits the possibility of error. Self-interpretation on the basis of globally broadcast information in addition to other stored information is therefore distinctly interpretive in character.

CONCLUSION

While it can be demonstrated that we sometimes self-ascribe by self-interpreting, scientific thought is not sufficiently advanced to determine the nature of all self-ascription empirically. The task of elucidating this process and the mechanism that enables it therefore falls to the philosopher. The task is an important one, in order to form effective experiments, scientists require a conceptual model (section 2.4), a framework of assumptions about the nature of the target domain, what Kuhn (1962) called a paradigm. Theoretical advances are required before empirical advances can be achieved. Philosophy can therefore be viewed as preparing the ground for science, performing the task of engineering the conceptual apparatus to be exploited by scientific examination. Philosophical analysis ensures that the way that scientists' think about a subject is both internally consistent and consistent with the evidence available. This essay presents the case for thinking of our knowledge of ourselves as interpretive rather than introspective in nature.

In the first chapter I demonstrated that the introspective intuition provides no support for an introspective account of self-ascription. Moreover, I showed that self-ascription can be achieved by a process of self interpretation. The only reason to still posit a separate introspective mechanism is the belief that self-interpretation is not a plausible account of *all* self-ascription. In order to counter this I then go on to develop an account of *theoretical self-interpretation* aimed at providing a plausible account of *all* self-ascription. In the second chapter I examine the nature of intentional states and establish that they are theoretical posits of TOM. In the third chapter I show how TOM enables us to perceive our intentional states through a process of theory-laden perception. With a plausible account of theoretical self-interpretation in place, introspection is stripped of all of its remaining plausibility.

Introspection is unable to explain certain instances of self-ascription (Chapter 1) and so is implausible as an account of *all* self-ascription. The theoretical self-interpretation account

established in chapters **2 & 3** therefore has greater explanatory power than the traditional account. Further to this in order to explain *all* instances of self ascription the introspectionist needs to resort to defending a dual method theory of self ascription which posits two separate ascription mechanisms in order to achieve the same explanatory power. Theoretical self-interpretation posits only a single mechanism and so is simpler than the dual method theory. Theoretical self-interpretation should be preferred for reasons of parsimony. The argument however is not decisive, while simpler theories are to be preferred all things being equal, this does not amount to irrefutable proof. Introspection is logically possible, but in this essay I have shown that the available information suggests that it is implausible in comparison to the theoretical self-interpretation defended here. Therefore until science provides us with more decisive proof, it is rational to favour the purely theoretical self-interpretation account over purely introspective theories (explanatory power) or dual method theories (parsimony).

BIBLIOGRAPHY

- Baars, B. J., Ramsay, T. & Laureys, S. (2003) 'Brain, consciousness and the observing self' *Trends in Neurosciences* **26** p671-675
- Botterill, G (1996) 'Folk Psychology and Theoretical Status' in *Theories of Theories of Mind*, eds Carruthers, P & Smith, P.K. Cambridge University Press. p 105-118
- Brasil-Neto, J.P., Pascual-Leone, A., Valls-Sole, J., Cohen, L.G., & Hallett, M (1992) 'Focal transcranial magnetic stimulation and response bias in a forced-choice task' *Journal of Neurology, Neurosurgery and Psychiatry* **55** p 964-966
- Carnap, R (2002 [1928]) 'The Logical Structure of the World and Pseudoproblems in Philosophy' Open Court Publishers: Colorado USA
- Carruthers, P (1996b) 'Simulation and Self Knowledge: A Defence of Theory-Theory' in *Theories of Theories of Mind*, eds Carruthers, P & Smith, P.K. Cambridge University Press. p 22-38
- Carruthers, P (2006) 'The Architecture of the Mind' Oxford University Press
- Carruthers, P (2008) 'Cartesian Epistemology: Is the theory of the self-transparent mind innate?' *Journal of Consciousness Studies*, **15**, No. 4, p 28-53
- Carruthers, P (2009) 'How we know our own minds: The relationship between mindreading and metacognition' *Behavioral and Brain Sciences*, **32**, p 121-182
- Churchland, P. M. (1981) 'Eliminative Materialism and the Propositional Attitudes' Reprinted in Lycan, W. G & Prinz, J. J. (eds.) (2008) 'Mind and Cognition' p 231-244
- Collin, F (1985) 'Theory and Understanding' Blackwell Publishing
- Dawkins, R (1976) 'The Selfish Gene' Oxford University Press
- Dehaene, S. & Naccache, S (2001) 'Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework' *Cognition* **79** p 1-37
- Frith, U & Happé, F (1999) 'Theory of Mind and Self Consciousness: What Is It Like to Be Autistic?' *Mind & Language*, Vol. 14, No. 1, p 1-22
- Frith, U & Frith, C. D. (2003) 'Development and Neurophysiology of Mentalizing' *Philosophical Transactions of the Royal Society London B*, **358**, p 459-473
- Gazzaniga, M. S. (1995) 'Consciousness and the cerebral hemispheres' in Gazzaniga, M. S. (ed) 'The cognitive neurosciences' MIT Press
- Gazzaniga, M.S.(2000) 'Cerebral Specialisation and Interhemispheric Communication: Does the Corpus Callosum Enable the Human Condition?' *Brain*, **123**, 1293-1326
- Giere, R.N. (1988) 'Explaining Science' The University of Chicago Press
- Giere, R.N. (1999) 'Using Models to Represent Reality' in Magnani, L. Nersessian, N.J & Thagard, P. (eds.) 'Model-Based Reasoning in Scientific Discovery' Kluwer Academic/Plenum Publishers
- Godfrey-Smith, P (2005) 'Folk Psychology as a Model' *Philosophers Imprint*, vol. 5 no. 6
- Goldman, A. I (1989) 'Interpretation Psychologised' *Mind and language*, **4**, p 161-185

- Goldman, A. I (1993) 'The Psychology of Folk Psychology', *Behavioral and Brain Sciences*, **16**, p 15-28
- Goldman, A. I (2004) 'Epistemology and the Evidential Status of Introspective Reports' *Journal of Consciousness Studies*, **11**, No. 7-8, pp 1-16
- Goldman, A. I. (2006) 'Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading' *Oxford University Press*
- Gopnik, A (1993) 'How we know our minds: The illusion of first person knowledge of intentionality' *Behavioral and Brain Sciences*, **16**, p 1-14
- Gopnik, A & Meltzoff, A. N (1997) 'Words, Thoughts and Theories' MIT Press
- Gordon, R. M. (1996) 'Radical simulationism' in *Theories of Theories of Mind*, eds Carruthers, P & Smith, P.K. Cambridge University Press. p 11-21
- Hernandez Cruz, J. L. (1998) 'Mindreading: Mental State Ascription and Cognitive Architecture' *Mind & Language*, Vol. 13, No. 3, pp 323-340
- Jastrow, J (1899) 'The mind's eye' *Popular Science Monthly*, 54, p 299-312
- Kosslyn, S. (1994) 'Image and brain' MIT Press
- Kuhn, T. S. (1996 [1962]) 'The Structure of Scientific Revolutions, Third Edition' *The University of Chicago Press*
- Langer, E (1975) 'The illusion of control' *Journal of Personality and Social Psychology*, 32 p311-328
- Lewis, D (1972) 'Psychophysical and Theoretical Identifications' *Australian Journal of Philosophy* Vol. 50, No. 3; December, 1972
- Lewis, D (1994) 'Reduction of Mind' in Guttenplan, S. (ed.) 'A Companion to the Philosophy of Mind' Blackwell p412-431
- Milgram, S. (1963) 'Behavioural study of obedience' *The Journal of Abnormal Psychology*, 67 p371-378
- Maibom, H (2003) 'The Mindreader and the Scientist' *Mind & Language*, Vol. 18 No. 3 p296-315
- Nichols, S. Stich, S. Leslie, A. & Klein, D. (1996) 'Varieties of Offline Simulation' in *Theories of Theories of Mind*, eds Carruthers, P & Smith, P.K. Cambridge University Press. P39-74
- Nichols, S & Stich, S.P. (2003) 'Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds' *Oxford University Press*
- Nisbett, R. E. & Wilson, T. D. (1977) 'Telling More than we can Know: Verbal Reports on Mental Processes' *Psychological Review*, Vol. 84 No. 3
- Quine, W.V. (1951) 'Two Dogmas of Empiricism' reprinted in Quine, W. V. (2004) 'Quintessence: Basic Readings from the Philosophy of W.V. Quine' *Harvard University Press*
- Quine, W.V. (1969) 'Epistemology Naturalised' reprinted in Quine, W. V. (2004) 'Quintessence: Basic Readings from the Philosophy of W.V. Quine' *Harvard University Press*
- Quine, W.V. (1996) 'Progress on two fronts' reprinted in Quine, W. V. (2004) 'Quintessence: Basic Readings from the Philosophy of W.V. Quine' *Harvard University Press*

- Robbins, P. (2004) 'Knowing Me, Knowing You: Theory of Mind and the Machinery of Introspection' *Journal of Consciousness Studies*, **11**, No. 7-8, pp 129-143
- Rosenthal, D. M. (2005) 'Consciousness and Mind' Oxford University Press
- Russell, B (1996 [1914]) 'Our Knowledge of the External World' Routledge: London
- Saxe, R (2005) 'Against Simulation: The Argument from Error' *TRENDS in Cognitive Sciences*, Vol. 9, No. 4
- Sellars, W, (1997 [1956]) 'Empiricism and the Philosophy of Mind' *Harvard University Press*
- Sellars, W (2007) [1954] 'Some reflections on language games" in Sharp, K and Brandom, R. B. 'In the space of reasons: Selected essays of Wilfrid Sellars' *Harvard University Press*
- Stanovich, K (2009) 'Distinguishing the Reflective, Algorithmic and Autonomous Minds: Is it Time for a Tri-process Theory?' In Evans, J. St. B T & Frankish, K. eds. (2009) 'In Two Minds: Dual Processes and Beyond' p55-88
- Stich, S. P. (1978) 'Beliefs and Subdoxastic States' *Philosophy of Science*, 45, p499-518
- Stich, S. P. (1996) 'Deconstructing the Mind' *Oxford University Press*
- Tversky, A. & Kahneman, D. (1974) 'Judgement under uncertainty: heuristics and biases' *Science* 185 p1124-1131
- Zahavi, D & Parnas, J. (2003) 'Conceptual Problems in Infantile Autism Research: Why Cognitive Science Needs Phenomenology' *Journal of Consciousness Studies*, **10**, No. 9-10, pp 53-71