ALIENS, DREAMS AND STRANGE MACHINES:

An Investigation into Thought, Interpretation and Rationality

Christina Cameron

Sidney Sussex College, University of Cambridge, 3rd April 2013

This dissertation is submitted for the degree of Doctor of Philosophy.

PREFACE

Declaration:

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. No part of this dissertation has been submitted for any other degree or qualification. It is approximately 78 000 words, including appendices and references.

Acknowledgements:

First of all, I would like to thank my supervisor Jane Heal, without whom this thesis would not exist. I am grateful for her willingness to read and reread the many drafts I have produced, for her enthusiasm in discussing aliens, and for her encouragement when I wanted to give up. She has helped me to work out what I think about thought, and then to express my ideas far more clearly than I could have done without her aid.

Thanks also to Simon Blackburn and to Tim Crane, my shadow supervisor, for all their help and advice.

Lucy Campbell, Chris Cowie, Alexander Greenberg and Lorna Finlayson each provided invaluable comments on parts of my thesis, and I have also benefitted from helpful discussions with Adrian Boutel, Tim Button, Hallvard Lillehammer, Steven Methven, Sebastian Nye, Adam Stewart-Wallace, Robert Trueman and Nathan Wildman.

Many of the above have also given me much non-philosophical help and encouragement, in which they were joined by Claire Benn, Amanda Cawston, Cley and Jane Crouch, Rachel Hilditch, Raymond Geuss, Lesley Lancaster, Basim Musallam, Remi Oriogun-Williams and Francis Young. This help has taken many forms, from free food and tea delivery, to help in proof-reading and computer aid, to impromptu counselling sessions. I am profoundly grateful to all of them, and would not have got to this stage without them.

Finally, I would like to thank my parents, for supporting me in so many ways throughout my education, and for always being there when I needed them.

This thesis was funded by the AHRC.

SUMMARY

Interpretationism about the mind claims that we can gain a philosophical understanding of the nature of thought by considering how we interpret the thoughts of others. My thesis aims to develop a version of this theory which is plausible in the sense that:

> (1) it has the potential to retain certain advantages attaching to theories of mind which focus on the behaviour, rather than the internal make-up of candidate thinkers;
> (2) it can fend off certain apparent counterexamples.

The thesis is split into four parts.

Part I explains why one might want to answer 'No' to the question 'Are there particular sorts of internal organisation which a being must have in order to count as a thinker?' It then introduces interpretationism as a position which will allow us to answer 'No' to this question. My version of interpretationism claims that a being has a thought iff it is interpretable as having that thought, and that all thinkers are rational. Both claims face several apparently obvious counter-examples. Parts II and III address these counterexamples by developing the crucial notions of interpretability and rationality.

Part II starts by considering the problem of seemingly hidden thoughts which occur during dreams, and uses this to develop an account according to which a subject is interpretable as having a thought if either a) there is sufficient evidence concerning the thought in the subject's actual situation and actions, or b) there would be sufficient evidence in at least one suitable counterfactual situation. I consider and reject an objection that this understanding of interpretability is incompatible with a commitment to the holism of interpretation, and then show how it can be used to address further proposed counter-examples, such as cases involving deception or paralysed thinkers. However, I agree with Block (1981) and Peacocke (1983) that their string-searching machine and Martian marionette must be counted as thinkers by this account. I argue that these are not counterexamples to the theory, however, because the intuitions against counting such beings as thinkers can be discredited.

Part III uses considerations about human limitations and propensities towards reasoning errors to argue that the interpretationist cannot adopt a deontological understanding of rationality that seems prevalent in the literature, nor a purely consequentialist account of rationality. I explain how Cherniak's (1986) conception of minimal rationality may be adapted for the interpretationist's purposes. I then consider and reject the idea that the emphasis on the rationality of thinkers will leave us unable to fit paradigmatically non-rational thoughts and thought processes (dream thoughts, imaginings and association) into our account.

Part IV shows why interpretationism so developed is well placed to retain the advantages of a theory of mind which focuses on behaviour, and outlines potential avenues for further research.

Contents

PART I: INTRODUCING INTERPRETATIONISM

Interpretationism is a position in the philosophy of mind which claims that we can gain a philosophical understanding of the nature of thought by considering how we interpret the thoughts of others. The purpose of this thesis is to argue that a kind of interpretationism is plausible in the sense that:

1. it has the potential to retain some advantages attaching to theories of mind which focus on the behaviour, rather than the internal make-up of candidate thinkers;
2. it can fend off certain apparent counterexamples.

Part I introduces the interpretationist approach. I begin, in chapter 1, by identifying a philosophical question to which interpretationism provides a response, distinguishing between two varieties of answer that have been given to this question, and then explaining some advantages of the type of answer which focuses the behaviour of thinkers.

Chapter 2 then argues that interpretationism provides just such an answer, while also improving upon analytic behaviourism. However, it admits that two notions which are central to interpretationism, *interpretability* and *rationality*, require further explanation, and that the theory faces several apparently obvious counterexamples. Part I therefore ends with an outline of how, in the rest of the thesis, I tackle these problems.

Chapter 1 – A Central Question

1. Visit to an alien earth

Imagine that some humans visit an Earth-like planet called Analog. On Analog, the explorers find life, and not only the simplest forms of life (creatures like viruses or amoeba in our own world); they discover creatures who are thinkers, who have intentional states such as beliefs and desires. Call these the Analogoids.

A group of humans, then, discover that Analogoids can think. How could this discovery be made? What sort of evidence would be relevant in making such a discovery? What evidence would be necessary for the humans to reasonably draw this conclusion? How certain could they be about their conclusion? And what consequences would the discovery have for how the explorers could then interact with the Analogoids?

I will use this thought experiment to guide an investigation into the nature of thinking. This will concern those mental states most uncontroversially taken to have intentional content: to be *about* something. This includes beliefs, desires, hopes, entertainings, imaginings and so forth. I call such states 'thoughts'.

My project, then, is to investigate what it takes to be a thinker, and this investigation will proceed via consideration of the following central question:

> Are there particular sorts of internal organisation which a being must have in order to count as a thinker?

Here, the word 'internal' indicates that the organisation of states, mechanisms etc. in question are supposed to be internal to the creature's thinking apparatus, and should not be behavioural events. It is not supposed to rule out that, even while

answering 'Yes' to the above question, we might hold that some being does its thinking outside of its body, for example because it thinks using a computer which is connected by radios to a robot body.

The phrase 'particular sorts' indicates that the question concerns whether our concept of thought includes (or should include – the idea that our concept may need revision should not be ruled out at this stage) a commitment to the idea that thinking involves certain kinds of organisation which must be described by talking about more than just the behaviour and possible behaviour of the thinker. Such an organisation and the states which constitute it will probably be taken to cause, or to have the potential to cause, certain behaviours. However, the person who answers 'yes' to the above question and then gives an account of what is needed in order for a being to be a thinker must, as Jackson and Pettit (1993) say, 'tak[e] on a substantial commitment to the nature of the underlying causes of behaviour.' (299) Some examples of such commitments are given in the next section.

If we answer 'Yes' to this question then the humans, when they say they have discovered that the Analogoids are thinkers, are saying that they have discovered that the Analogoids have such particular sorts of internal organisation (although this need not be inside anything recognisable as a head). Anything that is evidence of the existence of such inner stuff is relevant for the making of this discovery. The behaviour of the Analogoids may therefore be relevant as evidence of what is going on inside them. It may even be *very good* evidence, strong enough to warrant reasonable belief in the explorer's conclusion. However, no matter how much behavioural evidence the humans have, they need more information to guarantee that the Analogoids are thinkers. In particular, somehow looking more directly into the creatures' thinking apparatus and finding out what is happening in there is also

relevant – and on most accounts in the 'Yes' camp necessary – to guarantee the conclusion about Analogoid thinking.[1]

On the other hand, if we answer 'No' to this question, we presumably have to say that the behaviour of a creature is more than mere fallible evidence of whether that creature is thinking. We are then left with some further difficult questions to answer, about the nature of the relationship between behaviour and thought, about what behaviour is relevant to discovering whether or not a creature has a mind (or a particular thought), and about what that behaviour needs to be like in order for the creature to have a mind (or particular thought).

2. Two camps in the philosophy of mind

The central question of the previous section could be used to split philosophers of mind into two camps.[2] There is a strong tradition which says, 'Yes, there are particular sorts of internal organisation which a being must have in order to count as a thinker.'

There are many very different positions within this 'Yes' camp: the tradition includes substance dualists, who say that thinking must take place 'in' a special, non-physical substance; it includes the attempts of type-type identity theorists to identify types of mental states with types of brain states; and it includes at least some brands of functionalism, namely those which see the independently describable internal causal organisation of states of the brain as important. The important point

[1] See positions 2 and 3 at the end of the next section for examples of theories which answer 'Yes' to the central question but nevertheless would not require us to look inside a creature's thinking apparatus to guarantee that it was a thinker.

[2] Of course, these are only two camps into which we could divide such philosophers. There are other important and central questions which could be used to divide up the field in different ways, for example the question of whether statements about the mind can be reduced to statements in the language of the hard sciences, and the question of whether we should be realists or anti-realists about the mind. I believe it is possible to be a realist or an anti-realist and a reductionist or an anti-reductionist in both the camps I describe, but will not argue for this in this thesis.

of agreement among all these positions is that what is happening inside a creature's thinking apparatus is of utmost importance.

From the writings of contemporary philosophers, a paradigm example of such a theory is offered by Fodor, who claims that understanding the mental requires us to postulate internal representational systems which utilise a language of thought.[3] Searle is another 'Yes' camp member, stating that '*Epistemically*, we do learn about other people's conscious mental states *in part* from their behaviour… But *ontologically* speaking, the phenomena in question can exist completely and have all of their essential properties independent of any behavioural output.' (1992: 69) Block should also be counted as a member of the 'Yes' camp, due to his claim that 'whether behaviour is intelligent[4] behaviour depends on the character of the internal information processing that produces it.' (1981: 5) Block cautions against giving a positive characterisation of the sort of information processing required for thought, and argues only that certain types of information processing cannot involve thought, regardless of the behaviour they produce.[5] Still, he makes a commitment to the nature of the underlying cause of intelligent behaviour, and this is enough to place him in the 'Yes' camp. There are numerous other examples, which I will not list here.

There is, however, a second camp in the philosophy of mind, members of which answer, 'No, there are no particular sorts of internal organisation which a being must have in order to count as a thinker.' What unities the members of this camp, then, is the denial of the central claim of the 'Yes' camp.

One example of this tradition is Turing's approach to artificial intelligence. In his (1950) paper 'Computing Machinery and Intelligence', Turing proposed that we could replace the question 'Can machines think?' with 'Are there imaginable digital computers which would do well in the imitation game?'(442) The idea was that if a machine could fool a human conversational partner into thinking that it was

[3] See for example his (1976).
[4] Intelligence here refers to the capacity for thought or reason.
[5] See chapter 7.

another human in a text only conversation, then we could count that machine as intelligent. Turing took it that his test was overly stringent: he thought that computers might possess thought without being able to pass the imitation test. Nevertheless, passing the imitation test was supposed to be sufficient as a demonstration of intelligence: no matter what was going on inside the computer, if it could pass this test, it was a thinker.

Another philosopher who placed himself firmly in the 'No' camp was Wittgenstein, as reflected in the following comments from *Zettel:*

> 608. No supposition seems to me more natural than that there is no process in the brain correlated with associating or with thinking; so that it would be impossible to read off thought processes from brain processes. I mean this: if I talk or write there is, I assume, a system of impulses going out from my brain and correlated with my spoken or written thoughts. But why should the *system* continue further in the direction of the centre? Why should this order not proceed, so to speak, out of chaos?
>
> 609. It is thus perfectly possible that certain psychological phenomena *cannot* be investigated physiologically, because physiologically nothing corresponds to them. (1981: 106)
>
> Why should there not be a psychological regularity to which *no* physiological regularity corresponds? (Ibid.)

Wittgenstein's point here appears to be, not that he thinks that there couldn't be any causal regularity at the neural level, but that there *needn't* be. And presumably, since psychological phenomena could turn out not to correspond systematically to anything at the physiological level, something other than the kinds of internal states and mechanisms a creature has must determine which mental states the creature has.

Two more recent philosophers who have answered 'No' to the question are Donald Davidson and Daniel C. Dennett. Davidson, for example, says 'What a fully

informed interpreter could learn about what a speaker means is all there is to learn; the same goes for what the speaker believes.' (2001: 148) Given the context, and Davidson's philosophy as a whole, it is clear both that this fully informed interpreter is not supposed to have information about the inner workings of the subject's thinking apparatus, and that Davidson intends the claim to apply to more than just beliefs. Dennett, on the other hand, says that 'all there is to really and truly believing that p (for any proposition p) is being an intentional system for which p occurs as a belief in the best (most predictive) interpretation.' (1987: 29) Again, given the context, it is clear that coming up with an interpretation is not supposed to involve finding out about the inner workings of the creature's thinking apparatus, and that the claim is supposed to apply not just to beliefs, but to thoughts more generally. Indeed, Dennett says that his theory of mind is 'maximally neutral about the internal structures that accomplish the rational competences it presupposes'. (2009: 346) Davidson and Dennett's views are considered in more detail in chapter 2, and in later parts of this thesis.

One might object that the central question introduced above does not divide the philosophical landscape neatly in two, because it is possible for a philosopher to sit on the fence between the two camps. For example, a philosopher might say that certain behaviour is not merely a sign of certain inner states and processes, but a *guarantee* of a particular interior organisation, so that inner structure and outer behaviour cannot be separated in the way that both camps assume.

To answer this criticism, we must distinguish three of the ways in which a philosopher might link internal organisation to behaviour or the potential for behaviour, and to thought. These three positions are outlined below:

1) According to our concept of thought, what is really important about thought is the sort of behaviour that results or can result from it. Inner stuff is only relevant insofar as it has the potential to affect outer stuff. However, it just

so happens (for example, because of empirically discoverable laws of nature) that the behaviour that is important to thought can only result from a specific kind of inner stuff, and so we can know, given the relevant behaviour, what the inner stuff producing it is like too. Nevertheless, inner stuff is not a part of our concept of thought.[6]

2) According to our concept of thought, thinking has to involve particular kinds of internal organisation. However, it also just so happens that this inner stuff is the only thing which can produce behaviour of a particular kind, and so this sort of behaviour guarantees that the right sort of inner stuff is present. However, this behaviour is no part of our concept of thought.
3) Our concept of thought involves both the idea of particular sorts of internal organisation, and the potential for a certain sort of behaviour. Each also happens to guarantee the existence of the other.

I count the second two options above as positions within the 'Yes' camp: they involve saying that a being must, according to our (or a related and better) concept of thought, possess particular sorts of internal organisation in order to count as a thinker, even though they allow that (the possibility of) certain behaviours might be an infallible indication of thought, and (in the case of position 3) also necessary for thought. Only the first answer, which says that our concept of thought either does or perhaps should contain no commitment to any particular sort of inner stuff, counts as a position within the second camp. To be in the 'No' camp, one must answer that, for all that is contained in the appropriate concept of thought, a thinker does not need to possess any particular sort of internal organisation.

Given the assumption that we want to give the same account for all kinds of thought, I therefore conclude that the two camps I have distinguished within the

[6] For example, it might turn out that, in this world, only carbon based beings can produce behaviour complicated enough to warrant the attribution of thought. Still, a silicon-based thinker is not conceptually impossible.

philosophy of mind provide a reasonable way to divide the terrain, since given this assumption one can only enter the second camp by denying the central claim of the first camp. However, there is at least one possible position which still resists this sort of classification. This position is discussed in Appendix 1. For now, I proceed on the assumption that the central question I have identified is a useful question to ask.

Since there are many philosophers who fall into the 'Yes' camp, someone who answers 'No' to the central question is making a contestable and interesting claim. The next section explores some reasons why someone might want to make such a claim.

3. Reasons for answering 'No'

Recall that we are considering how to answer the following question:

> Are there particular sorts of internal organisation which a being must have in order to count as a thinker?

There are various reasons why one might answer 'No' to this question, and some of them have to do with long philosophical arguments and/or complicated constellations of other philosophical commitments. However, there are three quite intuitive considerations which may make a 'No' answer seem preferable. I will focus on these.

3.1 Knowledge of minds

If commonsense, everyday thinking about the mind involves a commitment to the existence of particular sorts of internal organisation, then we might worry that, given the lack of access the ordinary person has to such inner stuff, this leaves our beliefs about minds very vulnerable to sceptical attack.

In the first instance, this worry may present itself as a version of the problem of other minds: if when I suppose that you think x, I am supposing that an event y with such and such properties is taking place within your brain, but I have no evidence that this is occurring save for your behaviour, which by hypothesis cannot be enough to ensure the existence of event y, then how can I have sufficient justification for my belief? The standard answers to the problem of other minds may then be employed: we can suggest that we justifiably believe that other people have certain thoughts, or indeed any thoughts at all, by an analogy with our own case, or because it is the hypothesis that best explains their behaviour, just like unobservable entities are justifiably postulated in scientific theories.

However, if we adopt certain positions within the 'Yes' camp, namely those which identify mental events with certain independently identifiable physical states, events, etc. – or at least say that mental events must involve such states, event etc. – then there is also a more radical problem for us to face. If we say that when we think about minds we are speculating about such inner, physical stuff, then it seems that we may not even have unproblematic access to *our own* mental lives. This further threatens our knowledge of other minds: the argument from analogy cannot work if we do not have a case to make an analogy from. But it also suggests that we might be substantially mistaken about even our own mental lives.

The idea that we are radically mistaken about our own and other's thoughts is accepted by eliminativists about the mind, such as Paul and Patricia Churchland. Given the assumption that thought about the mind involves speculation about physical goings on inside a creature, they suggest the possibility that when we

develop the tools to look inside and attempt to gain independent confirmation of folk psychology (the things ordinary people believe about minds), we won't find what we are expecting. Indeed, the Churchlands argue that there are already reasons to doubt that our commonsense ideas about the mind constitute a good theory.[7]

Of course, there are also dualist theories within the 'Yes' camp. However, quite apart from seeming unappealing for other reasons, it is not clear that these can help us with this problem. As well as potentially failing to find the physical structures that physicalist Yes-campers posit, advancing science might also reveal that there is no place in the causal chain for an immaterial substance to cause our behaviour.[8]

We are left with the possibility that our mentalistic talk just fails to refer: that there aren't really any beliefs and desires or hopes and dreams, in ourselves or in other people. Another possibility is that although our mentalistic talk does refer, we have radically mistaken beliefs about the things it refers to: for example, perhaps we are wrong about what beliefs and desires influence our actions, perhaps wrong about the idea that beliefs aim at truth, or wrong about the idea that reasoning ever precedes actions, rather than merely providing post hoc justifications.

Now, one might reply to this sort of worry (as some have replied to the Churchlands) by saying that we have good reason to think that our commonsense ideas about the mind won't be undermined in this way. We might say that the truth of commonsense psychology is the best explanation of how useful we find it, how good it is at generating predictions, and so on. Such a reply agrees that commonsense psychology is a theory which postulates unobservable things whose existence we can and should be able to confirm in other ways. It just argues that it is such a good theory that we can be very confident about finding such independent confirmation, and are already quite justified in employing it.

---

[7] See for example Churchland (1989).

[8] Cf. Lewis (1966).

This response, however, concedes that all our most cherished beliefs about the mind are *vulnerable* to scientific refutation, even if such refutation looks unlikely. We might think that this already concedes too much, that we can already be certain that we know some things about minds.

A line of thought that at least some will find intuitive, then, says that there is something special about minds and our knowledge of them in comparison to other unobservable things dealt with by scientific theorising. According to this line of thought, we have to be substantially right about at least some important aspects of thought. We do not need to worry about turning out to be drastically wrong about these things, not just because it's unlikely, but because these ideas are not vulnerable to refutation in the same way that some of our other beliefs are. This line of thought might also suggest that we do not need to show that commonsense thinking about the mind is good in the same way that some scientific theories are good.

It would, of course, be possible to push this line of thought too far. We do not want to deny that there are many similarities between the ways we interact with and find out about other people and the ways we interact with and find out about other sorts of things. We also shouldn't deny that our beliefs about minds, even our own minds, are fallible. We should accept that we can fail to understand other people or misjudge their motives and feelings, and that we sometimes engage in self-deception or display an unfortunate lack of self-knowledge. Still, this does not mean that there is nothing to the idea that our beliefs about the mind have some immunity from scientific refutation. And the inhabitants of the 'No' camp seem to face an easier task in explaining what is right about this idea. If being a thinker consists in being disposed to or having the capacity to behave in certain sorts of ways given certain circumstances, then in a lot of cases we already have the evidence that we need to be sure of the existence of (and important aspects of the nature of) minds. And if what goes on inside a person doesn't matter as long as they (could) display the right behaviour, then we can be totally relaxed in our theory of mind regardless of what scientists may or may not find inside our heads.

This isn't supposed to be a conclusive argument, and there are of course things that members of the 'Yes' camp can say about our knowledge of minds. Nevertheless, this offers us a reason why, at least at first glance, the 'No' camp might look like an attractive place to be.

3.2 Avoiding chauvinism

Suppose that humans do, in fact, have the right sort of inner stuff to count as thinkers according to your chosen 'Yes' camp theory. In that case, our claims to knowledge of human minds might be a little shaky, but at least our beliefs about our own and other humans' mental states could be largely true. Even so, the 'Yes' camp theory might still get the extension of 'thinker' wrong.

Return to the example of the Analogoids. Suppose that the explorers meet them, make contact, develop a way of communicating with them, enter into trade relations with them, enjoy social occasions with them, and believe that they have developed relationships of trust and friendship with them (a belief with which the Analogoids seem to claim to concur). Then one day, a human scientist has an opportunity to look inside an Analogoid's head-equivalent, and discovers that what is going on in there is totally different from what goes on inside a human brain, in such a way that the Analogoids do not fulfil your chosen 'Yes' camp theory's conditions for being a thinker. Suppose, in fact, that Wittgenstein's hypothesis holds true of the Analogoids: there seems to be no inner physiological regularity correlated with their outer behaviour at all. I take it that this would mean that the Analogoids would not fulfil the conditions for being a thinker on *any* materialist 'Yes' camp theory.

It seems as though, if you have answered 'Yes' to the central question above, and you reject dualism, you ought to say that the scientist has discovered that the party of explorers were wrong about the existence of Analogoid thought: and

therefore wrong that what they had developed with them was a system of communication, and wrong that they had built relationships with them. And if you accept that conclusion, it seems as if this should have a knock-on effect on how the humans should then treat the Analogoids: for example, perhaps it should mean that they are not morally obligated to honour trade agreements.

But this seems both implausible and very unfair. Surely scientific evidence about what goes on inside the Analogoids shouldn't be able to undermine the significance of the Analogoids' ability to interact with us? This result suggests that theories in the 'Yes' camp are what Block calls 'chauvinist' (Block defines chauvinism by saying that 'theories are chauvinist insofar as they falsely *deny* that systems have mental properties.' (1980b: 292))

An easy way to avoid the problem would be to move into the 'No' camp. If the explorers' conclusions about Analogoid thought weren't conclusions about the existence of particular sorts of inner states, mechanisms, structures (etc.) in the first place, then the discovery that such inner stuff does not exist wouldn't threaten the conclusion about thought.

Again, there may be ways to try to avoid this problem from within the 'Yes' camp. However, for the moment we can take this as another prima facie advantage of answering 'No': it seems to make it easier to avoid chauvinism.

3.3 Origins and uses of psychological concepts

Ordinary conversation is replete with psychological concepts: we talk about what we and other people think and want, what we intend to do, how we feel about various situations, how sure we are about the things we believe or want to do. We are taught to use such language as children, and the relevant terms ('think', 'want', 'hope', 'intend', 'fear', etc.) are introduced to us in the context of certain behaviours

in certain environments. This is true both of being taught to apply such words to other people, and being taught to apply them to ourselves.

In the case of applying such terms to other people, the confirmation and disconfirmation conditions we use in everyday life for the application of psychological concepts involve behaviour, not the inner workings of those other people's brains. This is so simply because we do not ordinarily have any access to the inner workings of their brains, except insofar as this results in behaviour. Our only option is to watch what people do and listen to what they say. In our own case, things seem to be somewhat different. It appears that we don't usually need to look at our own behaviour in order to discover what we think. However, even if we do have a special sort of access to our own thoughts, it is difficult to explain how this can play a role in fixing the content of shared concepts, expressed by words in a shared language, except insofar as it results in behaviour.

Given certain theories of meaning, most notably verificationist theories, these sorts of considerations might be enough to establish that psychological concepts concern behaviour, rather than hidden inner states or mechanisms. However, one does not have to be a verificationist in order to think that the circumstances in which a concept is introduced and learned are relevant to its meaning, and one might further think that reasons need to be given for thinking that the meaning goes beyond such considerations in a particular case.

Such a reason might be found in the use to which we put particular concepts. But in the case of psychological concepts, when we look at the role they play in our lives, we might think that we find further reason to suppose that these concepts do not need to concern inner stuff. Many philosophers have emphasised the role such concepts play in explaining, predicting and manipulating the *behaviour* of other people.[9] Others, for example Heal (2003), suggest that there is something wrong with taking our dealings with each other to be so similar to the ways in which we

[9] Cf. Dennett (1987) and Cherniak (1986).

explain, predict and manipulate ordinary physical objects. Still, they emphasise the role of psychological concepts in structuring our interactions with each other: they just think that a broader understanding of these interactions is needed.[10]

Support for the idea that this is indeed how we use psychological concepts, and that this use should incline us towards positions in the 'No' camp, can again be found by looking at our reaction to the Analogoid case in the previous section. It seems plausible that it is the fact that psychological concepts provide such satisfying explanations of Analogoid behaviour, and facilitate fruitful interaction, that makes it seem so reasonable to apply the concepts to them and so unreasonable to withdraw attributions of thought, regardless of what is going on inside.

Once again, I don't pretend to have offered anything like a knock-down argument here. Rather, I present these thoughts as another source of prima facie support for a 'No' answer, suggesting merely that certain facts about the origins and uses of psychological concepts might be easier to account for if we situate ourselves in the 'No' camp.

4. Conclusion

I conclude that there is some intuitive plausibility to the claim that is definitive of what I have called the 'No' camp – i.e. the claim that our concept of thought does or perhaps should not entail that there are particular sorts of internal organisation that a creature must have in order to count as a thinker. If we want to give this answer, however, some difficult questions arise, as outlined in section 1: questions about the relationship between behaviour and mind; about what sort of behaviour is relevant to whether or not a creature has a particular thought; and about what that behaviour needs to be like in order for the creature to have a mind or particular thought. In

[10] This issue will be discussed further in chapter 10, section 6.

order to capitalise on the suggested advantages of being in the 'No' camp, we need to show that something can be said in response to these questions.

In the next chapter, I introduce interpretationism as a position in the 'No' camp.

Chapter 2 - Interpretationism

The purpose of this chapter is to characterise interpretationism in enough detail that it becomes clear:

1. What its central features are;
2. How it differs from another, more famous, theory in the 'No' camp, namely analytic behaviourism;
3. What varieties of interpretationism are on offer; and
4. How we need to develop interpretationism if we are to produce a clear and plausible theory.

I present the central features of interpretationism in sections 1 and 2. Section 1 characterises the position very roughly through two positive and two negative claims, and section 2 shows why the two positive claims come together through a consideration of the kind of interpretation that interpretationism takes to be so important in the philosophy of mind. Together, these sections illustrate the difference between interpretationism and analytic behaviourism, understood as the thesis that statements containing mentalistic terminology can be paraphrased or translated into statements which only use behavioural terminology.

However, at the end of section 2 important questions remain about the precise content of interpretationism's central claims and the kind of account of thought that interpretationism is trying to give. In section 3 I show that there is no one answer to such questions, because interpretationism as I have characterised it can be broken down into four sub-categories. One of these also exhibits a further difference from analytic behaviourism.

Finally, in section 4 I introduce some apparently obvious counterexamples to interpretationism, and outline my plan for tackling them and developing interpretationism further through the rest of this thesis.

1. Four claims

The position that I am interested in when I use the name 'interpretationism' is a position within the 'No' camp. It therefore denies the central claim of the 'Yes' camp, and so says that our concept of thought either does not or should not entail that there are particular sorts of internal organisation that a being must have in order to count as a thinker. It also makes four more distinctive claims, as I outline in this section.

The first positive claim is that a being has a thought iff it is interpretable as having that thought. I will call this the Availability Claim. There is much to be said about what interpretation and interpretability amount to, and this is discussed in some detail in section 2 below and throughout this thesis. In Part II, I distinguish different kinds of interpretability, and argue that the interpretationist should adopt a weak notion of it. However, for the purposes of outlining the central features of interpretationism, it is only necessary to say that the information required for the sort of interpretation in question concerns only the being's environment, its interactions with that environment and possibly some other behaviour. It does not concern the things happening inside the being's thinking apparatus.[11] Characterised in this sparse way, the Availability Claim is something that the analytic behaviourist agrees with.

[11] Unless the being in question has some manner of access to the stuff happening within its thinking apparatus using those of its senses which register things about its environment. Suppose, for example, that a creature had a transparent head, and that parts of its brain lit up when active. Suppose that this creature were to stand in front of a mirror watching a part of its brain that was alight and wondering whether the wondering it was engaged in had a spatial location within the glowing lobe. Aspects of the inner workings of the creature's thinking apparatus would then count as a part of the creature's environment and so some information about them should be allowed in determining whether their thoughts were interpretable.

It should be noted that the insistence on the sufficiency of interpretability is the major difference between Child's (1994) characterisation of interpretationism and my own: Child, being concerned with *any* view in which considering interpretation plays a central role in our understanding of thought, is also interested in those views on which interpretation of behaviour must be supplemented with other information, including information about the inner states, processes etc. that are picked out as important by people in the 'Yes' camp. What I simply call 'interpretationism', Child refers to as 'pure interpretationism'. (1994: 40) The claim about the necessity of interpretability, on the other hand, is important to my kind of interpretationism for the same reason that it is important to Child: because it allows us to say that interpretationism tells us something positive about the nature of thought.[12]

The second positive claim that interpretationism makes is that all thinkers are, in some sense, rational. I will call this the Rationality Claim. This is not something that analytic behaviourism needs to commit itself to, and indeed behaviourism has not generally been taken to make any claims about the rationality of thinkers. Nevertheless, a commitment to rationality could easily be added to the traditional analytic behaviourist theory. The Rationality Claim does not, therefore, immediately mark an important difference between interpretationism and analytic behaviourism.

The most important differences between analytic behaviourism and interpretationism are contained within two negative claims that I take interpretationism to make. The first of these involves interpretationism's answer to a question, posed in chapter 1, for anyone in the 'No' camp: what is the nature of the relationship between behaviour and mind? The analytic behaviourist says that the relationship is one of identity, whereas interpretationism, as I define it, denies this.

What then does the interpretationist say that thoughts are? I take it that different interpretationists might give different answers to this, but that one of the answers open to them is to say that individual thoughts (such as my current belief

[12] Cf. Child (1994: 23-4, 27 and 31-32).

that it is raining) are identical to inner states, events etc. (such as some particular current state of my brain). Interpretationism does not thereby pass into the 'Yes' camp, however, because it says that all that is important about such states, insofar as they count as thoughts, is what environmental factors produce them, what behaviours they do or could result in, and how the subject can therefore be interpreted. Thus, interpretationism makes no substantial commitments concerning the nature of the inner states, beyond their connection to the subject's surroundings and behaviour. If the interpretationist can make this proposal work, then they avoid an objection that is sometimes[13] levelled at analytic behaviourism: that it wrongly denies that our thoughts cause our actions. The interpretationist will be able to allow that thoughts cause the behaviours that we commonly suppose them to.

The second negative claim of interpretationism involves denying that there is anything useful to say about the behaviour connected to a particular thought (such as a desire for ice cream) or even a kind of thought (such as desire). Here again, interpretationism directly denies one of the claims of analytic behaviourism. Unlike behaviourism, interpretationism demands only that each thought be interpretable, and so it can accept that there could be many, even an infinite number, of total situated behavioural states which would allow an interpreter to interpret one particular thought (such as the desire for an ice cream). Of course, given each total situated behavioural state, the interpreter also needs to interpret many other thoughts as well, as I discuss further in the next section.

Since interpretationism does not try to analyse our talk about thought into talk about behaviour, this makes it look much less vulnerable to another objection that is sometimes levelled at analytic behaviourism: the Chisholm-Geach objection.[14] This objection accuses analytic behaviourism of circularity by arguing that for any particular thought or kind of thought, we cannot link it to behaviours or potentials for behaviour without talking about other mental states or kinds of mental state.

[13] See, for example, Braddon-Mitchell and Jackson (2007).
[14] See for example Block (1981).

Interpretationism cannot suffer from the kind of circularity attributed to behaviourism if it does not give the analyses in which this circularity was uncovered.

Interpretationism, then, is a position which says that thinkers are fully interpretable and in some sense rational, and which explicitly rejects two of the commitments of analytic behaviourism which triggered problems for that theory.

However, even while admitting that interpretationism is an alternative to *analytic* behaviourism, some philosophers would argue that it remains a brand of behaviourism. For example, Block (1981) and Davies (1991) appear to class any position within the 'No' camp as a kind of behaviourism. Davies would count a position like my version of interpretationism as a kind of 'supervenient behaviourism' (255), and based on the comments in his (1998) paper, I take it that Alex Byrne would agree.

I do not disagree that in some ways 'supervenient behaviourism' is an appropriate way to categorise interpretationism as I have characterised it. However, given the unpopularity of traditional behaviourism, it would be an off-putting label for many, and would perhaps discourage people from giving interpretationism the further consideration I think it deserves. Therefore, I will not be using that label.

So far, I have only explained interpretationism in very rough terms. To understand it better, and to show how its claims are connected and why it exhibits some of the differences from analytic behaviourism that it does, we must look more closely at the kind of interpretation that it takes to be important. This is the task of the next section.

2. The interpretation in interpretationism

In this thesis, and in the work of philosophers I count as interpretationists, 'interpret' and its cognates are used extensively. There is talk of the process of interpretation,

the method and assumptions of interpretation; of interpretation theory; of homely, radical, real and idealised interpreters; of people and languages being interpreted; of individual thoughts, actions or sentences being interpreted; and of 'interpretability', and what that could mean.

As Simon Blackburn has pointed out, the use of 'interpret' seems more at home in some of these contexts than in others. Blackburn also gives an example of a request for interpretation sounding odd: 'suppose I show you a familiar domestic object which is obviously, say, a spoon, and ask you 'James thought this was a spoon. How do you interpret that?' Then you would be at a loss. What are you supposed to say (absent some strange context)?'[15] In response to these comments, I will outline four possible meanings of 'interpretation' and explain how I take these to relate to the project of interpretationism.

First of all, 'interpretation' may be used as more or less synonymous with 'identification'. When the interpretationist talks about interpreting a thought or sentence, he is often referring to identifying that thought or sentence in the sense of establishing its existence and determining its content. The question of whether a thought or sentence is interpretable is then the question of whether it can be identified in this way using the method which the interpretationist considers to be important.

On the other hand, 'interpretation' may be used to refer to gaining a particular sort of understanding: the sort of understanding given by explanations in terms of reasons. For example, I might wonder how, in this sense, to interpret a ball being thrown in my direction: as a mistake, an attack, or an invitation to play.

'Interpretation' may also be used to refer to simplifying and/or clarifying something. For example, one might ask for an interpretation of a paragraph of Kant's *Critique of Pure Reason*, even in its English translation. One could thereby ask

[15] Personal correspondence.

for a simpler restatement of the paragraph, or a summary of its important points, rather than an account of the reasons why Kant wrote it.

Finally, 'interpretation' might refer to a process of attributing significance to an event or object, without explaining this by calling on the reasons of a person. So, for example, a doctor might be taken to interpret the symptoms of their patient, or a scientist might interpret the results of their experiment. This may also be the sort of interpretation applied, less successfully, to tea leaves, entrails, or configurations of stars in attempts at divination.

These four kinds of interpretation are related to each other. The second and the fourth both involve giving a kind of explanation of an object or event, but I think they are worth distinguishing because only the first has to involve attributing reasons to a particular being. The third may seem separate from the second as far as the questioner is concerned, but the person who wishes to give the third kind of interpretation of a paragraph of Kant probably needs to engage in the first two kinds of interpretation in order to do this well.

Interpretationism is concerned primarily with the first two kinds of interpretation. It is interested in a process of trying to identify thoughts through a process of understanding and explaining persons and their actions. The fact that the word 'interpret' rather than 'identify' is applied to thoughts reflects the fact that it is this particular method of identification that the interpretationist is interested in.

Given these considerations, we can partly explain why the word 'interpret' sometimes seems odd when applied to thoughts. Take Blackburn's example above: here we can say that the request to interpret is indeed misplaced, because the first sort of interpretation has already occurred, the second and third are unnecessary (the thought is stated as simply and clearly as one could wish, and there are no special explanations required for you to understand why a person might think that a spoon is a spoon), and the fourth sort of interpretation is not something we are used to applying to everyday thoughts about spoons. That there are different kinds of

interpretation, together with conversational implicature, creates much of the strangeness here.

This may not completely dissipate the oddness, however. One might object that 'interpretation' is far more naturally applied to the understanding of a language, and that there is an important disanalogy between language and the mind: in interpreting parts of a language, an interpreter can focus on what he is interpreting, a series of noises perhaps, or some marks on a page, in a way that we do not ordinarily think possible in the supposed interpretation of thought. Unless we are to identify thoughts with behaviour, as the behaviourist did, we cannot see or hear other people's thoughts. Thus one might object that they are not available to be interpreted until the identification has taken place. It may therefore continue to feel slightly inappropriate to talk about interpreting thoughts in the sense that the interpretationist intends.

This is perhaps one of the reasons why Child (1994) insists that a person, rather than their thoughts, is always the primary object of interpretation. Another reason for saying this relates to the impossibility of 'interpreting' thoughts one at a time, and the need to consider all of a person's interactions with their environment. This will be discussed in more detail shortly. I agree with Child in this matter. However, since I am interested particularly in thoughts which are difficult for an interpreter to identify and attribute, I will sometimes talk about what it means for particular thoughts to be 'interpretable' as a useful shorthand.

Having determined something further about what sort of interpretation interpretationism is interested in, we can add to our characterisation of the theory: we can say that interpretationism takes thoughts to be things which are identifiable by a process of understanding and explaining thinkers and their actions in terms of reasons, that thoughts are essentially involved in a creature acting for reasons, and that thinkers are beings susceptible to understanding in terms of reasons, and therefore creatures with purposes and reasons for doing things.

The Availability Claim, understood using this notion of interpretation, ceases to be something that the analytic behaviourist needs to commit to: the analytic behaviourist does not *need* to say that thinkers and their actions are always susceptible to explanation in terms of reasons in order to count as an analytic behaviourist, just as he does not need to make any commitments concerning the rationality of thinkers. The two are obviously connected, and this offers the beginnings of an account of why the interpretationist's Availability and Rationality Claims come together: it seems that a thinker needs to be rational to at least some degree in order for the kind of interpretation of them which is supposed to be possible to take place.

How exactly does the process of identifying thoughts through understanding and explaining thinkers and their actions in terms of reasons work? And what relationship does this sort of interpretation bear to our own real life methods of understanding each other? These are questions to which different interpretationists may give different answers. For example, different answers are given by the two most famous interpretationists, Dennett and Davidson.[16] Rather than engage in exegesis, I will briefly outline two common features of the two accounts to produce a sketch of how interpretation may proceed, and illustrate them using two examples. I will then comment on some of the differences between Dennett and Davidson's accounts.

According to both Davidson and Dennett, interpretation must begin by assuming that the subject is, at least to some degree, rational. It then proceeds by the interpreter working out, using this assumption together with *a lot* of information

[16] A clear account of the procedure Dennett envisages is given in his paper 'True Believers' (contained in his (1987)). Davidson describes the process of interpretation that concerns him in a variety of ways over several decades, and one may question to what extent Davidson's writings present us with a single, gradually elaborated picture of interpretation, and to what extent he changes his account over the course of his career. A good review is given in chapters 12 and 14 of Lepore and Ludwig (2005).

about the subject's history, environment and behaviour, what the subject must think at different times and, if they have language, also what they mean by their utterances. The interpreter must attribute *many* thoughts to their subject, and these must make sense not only in light of aspects of the subject's situation and behaviour, but also in light of each other: the thoughts attributed need to fit together.

First, take the example above of a person throwing a ball in my direction. In order to interpret their thoughts at the time, we proceed on the supposition that they had a reason for throwing the ball, and we try to work out what the reason might be. There are several candidates: they want me to participate in a game, they want to annoy me, they want to get my attention, and so on. To work out which is the real reason, we need more information.

We look, then, at other aspects of the situation. Suppose that this person and I are both standing near a basketball court. That makes the first option look sensible. We might refine our attribution, and say that he wants me to play basketball. However, to confirm this hypothesis, we must be sure of certain other things. For example, does this person have any experience of basketball courts and what happens on them? If he has never seen or heard of one before, then how could he desire to play basketball? But suppose we have seen this person on basketball courts quite often before, and he has often behaved, in conjunction with other people behaving similarly, in accordance with the rules of basketball. We might then attribute to him knowledge of the rules of basketball, and suppose that he frequently has the desire to play it.

However, given just this behaviour, there is still a lot we don't know about his frequent desire to play, and this may affect whether it makes sense to think he wants to play with me on this occasion. Does this person just love playing basketball? Or does he see it as a means to something else he desires? Basketball could be a means to many things: fitness, companionship, social status, his father's approval. We need to look at more of this person's situation and behaviour to find out the nature of his frequent desire to play.

Suppose, first, we look at his relationship with his father. This will, presumably, involve looking at quite a lot of linguistic behaviour on the part of both father and son, and so to make use of this source of information we need to be able to understand the language (or languages, or idiolects) in which father and son speak to each other. Suppose that we have enough information, and have it organised perspicuously enough, to be confident that we can understand what father and son say to each other, and suppose that we hear the father frequently complaining that the son spends too much time playing sports when he ought to spend more time studying. Perhaps the father suggests that he gives up basketball. That counts against the hypothesis that the son wants to play basketball to please his father. Suppose, moreover, that the son refuses, and gives as a reason that people like him because he's on the basketball team. That counts in favour of the hypothesis that at least part of the reason he wants to play basketball is because of the social status it gives him. We may then look at a lot of other verbal and non-verbal behaviour of this person, and consider potential reasons behind it, in order to further support the hypothesis that social status is very important to him.

Suppose the hypothesis about a desire for social status explains much of this person's behaviour very well. This may then have an effect on what reason we should attribute to him in throwing the ball at me. Suppose that I have very little social status. In conjunction with other behaviour from our peers, this may make it reasonable to suppose that playing basketball with me would damage this person's social status. And the person who threw the ball may have done many things in the past which indicate that he knows my social status, and knows how, according to the prevailing social norms, he therefore ought to treat me. Suppose also that there are other people around at the time that the ball is thrown, whose situated behaviour suggests that they are very keen enforcers of the social norms just mentioned, and suppose that the person who threw the ball at me has just uttered the word 'Hello,' while facing these people. We can then attribute to him the belief that they are there, and may be watching his behaviour.

Given all of this information, the hypothesis that the person wants to play with me seems very unlikely. It seems more likely that he wants to annoy me. Perhaps more information is required to confirm this. Still, it is already clear that in order to work out the reason behind quite a simple piece of behaviour, we have had to look at a great deal more behaviour than just the throwing of the ball, and have had to attribute many more beliefs, desires and intentions to the person who threw it.

We could also imagine this process of interpretation being used in the example from the beginning of chapter 1, the case of the humans who meet the Analogoids. Suppose that when the humans land on Analog, a group of Analogoids approach them as they exit their spaceship, one alien makes a series of noises, and then all of the aliens spread their arms outwards and up a little above their heads and display their teeth. The humans begin by supposing that the Analogoids wanted to do these things, that they had reasons for doing them. But based on such a small snapshot of Analogoid behaviour they do not know what the noises or the gestures signify about the Analogoids' thoughts. Of particular concern to the humans is the fact that they do not know what the Analogoids think about their arrival. Again, there are several options: the Analogoids might want to welcome them to their planet, they might want to warn the humans off, they might want to deliver a non-committal sort of greeting and to make their minds up about whether humans are good to have around later, or they might be in the middle of performing a religious ceremony which they had long planned to perform at this time and place, and which they feel compelled to continue with despite the unfortunate arrival of some odd-looking creatures in the relevant location. Suppose that the leader of the humans decides the best thing to do is to imitate the Analogoids: she makes a speech, then spreads out her arms and smiles, and her crew follow her lead. To everyone's relief, the Analogoids do not pounce on them, and the humans decide that the noises made by the lead Analogoid did not mean, 'Get back into that spaceship this instant or we shall rip you to pieces with our hands and teeth.' The humans and Analogoids then begin to observe and interact with each other.

The interpretation of the Analogoids will be more difficult than the interpretation of the ball-thrower, because the humans cannot call on shared knowledge of social norms etc. Additionally, interpreting the Analogoid's language is a very important part of the process and (unlike in the normal situations we encounter) it must start from scratch. These features make the interpretation of the Analogoids a case of what Davidson calls *radical interpretation*. This is the sort of case that he focuses on, and it makes particularly clear how much information about behaviour must be gathered and systematised, and why assuming that a subject has reasons for the things they do is necessary if you want to work out what they think. These features also, however, make it harder to describe a particular case concisely. The interpretation of language, for example, surely involves information about its use in all sorts of different situations, together with assumptions that the speakers are saying true and pertinent things with it on many occasions. This emphasises the fact that information about a great deal of behaviour is required, while also entailing that it would be very tedious to describe in detail how the humans work out what the first noises an Analogoid directed towards them meant.

The two examples above provide rough illustrations of the sort of process that both Dennett and Davidson have in mind when they talk about interpretation. However, as mentioned above, there are important differences between the two accounts. Two major differences include the centrality of prediction (Dennett emphasises prediction far more than does Davidson) and the importance of language (Davidson describes a process which centrally involves the interpretation of a subject's utterances, and argues that non-linguistic creatures cannot think. Dennett applies his method of interpretation to non-linguistic creatures, including very simple beings such as thermostats and clams). For the moment, I remain neutral on these issues. However, both will be discussed in Part III.

Dennett and Davidson also disagree on how the process they describe relates to what we actually do when we understand one another. Dennett thinks that the

process he describes is *our* process (1987: 11) and moreover that it is something we can't avoid doing with respect to ourselves and our fellows (1987: 27). On the other hand Davidson, at least in later work,[17] suggests that we cannot use his 'official' method of discovering other people's thoughts (radical interpretation), but that it reveals to us what the possibility of our interpretation of persons depends on. This is an important difference, because it affects the sorts of objections we might present. If we are told that the process of interpretation that is important is just *our* process, then we face two problems. First, we may worry that focusing too closely on our own interpretational abilities and methods, will, because of contingent human limitations, prevent us from being able to attribute thought to some possible thinkers. Second, what method we use to understand each other seems to be an empirical question. We may then worry that scientific discoveries about how our process of interpretation works may show that we don't use anything like the method that Dennett, for example, suggests. However, if we are told that interpretationism does not focus on our method of understanding each other, then it is much less clear how we can justify our approach by pointing to the way real humans acquire and use psychological concepts, and what they think they know about minds.[18]

There are, then, two potentially problematic ways to understand the relationship between the process of interpretation described by interpretationism and our actual methods of finding out about each others' minds. Perhaps the ideal solution would be to adopt some sort of middle position, which would allow us to avoid both sets of problems. However, in this thesis I will not attempt to give a full account of how this could be done. Rather, I keep the potential problems above in view, and aim to develop interpretationism in such a way that this issue will be tractable.

[17] See for example his (2004: 128).

[18] For this as a criticism of Davidson's approach, see Rescorla (forthcoming), which asks of Davidson's strategy 'Why this make believe?'

This means that the process of interpretation invoked must be at least closely related to our methods. It might be a description of our methods at a high level of abstraction, or it might be a rational reconstruction of our methods. It may in some ways be an idealisation of our method, but it is important that it not ignore important features of our situation that interpretationism is supposed to help us to understand. In particular, it must not ignore human limitations in such a way as to make it obscure how finite creatures such as ourselves can have thoughts and can interpret the thoughts of others. Whichever of the above best describes the relationship, it must be plausible to say that considering the method of interpretation picked out by interpretationism tells us what the possibility and reliability of *our* method relies on. I will comment on how this issue may be pursued further in Part IV.

At this point, I hope that the reader feels that they have some grip on what I mean by interpretationism. We do, after all, know something about what sort of interpretation it takes to be important, we know some of the claims that interpretationism makes about thinkers and thoughts, and we know some of the things that make interpretationism different from analytic behaviourism.

Still, there are some very important things we do not know. Take the claims that interpretationism makes about thoughts and thinkers: it says that thoughts are things which are identifiable by a process of coming to understand and explaining thinkers and their actions in terms of reasons and that they are essentially involved in a subject acting for reasons, and it says that thinkers are beings susceptible to understanding in terms of their reasons, and therefore creatures with purposes and reasons for doing things. But what is the status of these claims? What are they trying to do – to give us definitions of 'thought' and 'thinker', to provide an analysis of our concepts? Or are they supposed to do something else? Without knowing the answers to such questions, we cannot know what sort of a theory of mind we are dealing with, nor assess it appropriately. I have not answered these questions up

front because the sort of interpretationism I am interested in encompasses sub-types which give different answers to these questions.

In the next section, I introduce a four-way distinction between types of interpretationism, and explain the kind of account each attempts to give of thought, how each treats the central claims that any interpretationist makes about thought and thinkers, how they see the relationship between the Availability and Rationality Claims, and what particular problems each might face. As mentioned in the introduction above, one type of interpretationism also exhibits a further difference from analytic behaviourism.

3. Varieties of interpretationism

3.1 Derivative interpretationism

The aim of derivative interpretationism is to give an *analysis* of thought in independent terms. These independent terms do not include 'interpretation'. Nevertheless, according to derivative interpretationism it is a consequence of the correct analysis that every thought is interpretable. This seems to be the position that Child (1994) has in mind when he uses the label 'non-constitutive interpretationism'.

On this view, thought consists in a sort of pattern[19] in the life of one subject which can be specified without invoking the idea of interpretation. Two of the claims above: that thoughts are things which are identifiable by a process of coming to understand and explaining thinkers and their actions in terms of reasons, and that thinkers are beings susceptible to this kind of understanding, are then just the results of the analysis of thought and thinkers. They do not provide a definition or an analysis.

[19] Cf. Dennett's paper 'Real Patterns' in his (1998). I use the term very broadly, to cover both the actions of the creature over time, and its changing dispositional profile.

We must then ask what analysis the derivative interpretationist *will* give. How will they describe the patterns necessary for thought and thinkers? Well, these patterns will be ones which make the thinkers rational. I will take it that the derivative interpretationist will explain these patterns in terms of some notion of rationality, and that they will therefore take the Rationality Claim from section 1 as prior to the Availability Claim. It should be noted straight away, however, that the notion of rationality in play here may not be the notion that immediately comes to mind when we hear the word 'rationality'.

In some ways, derivative interpretationism is an ambitious theory: it aims, after all, to give a proper analysis of the concept of thought, and it may even ultimately try to analyse psychological concepts in terms of non-psychological ones (if rationality is not taken as a psychological concept, or can be further analysed in terms of non-psychological concepts), and intentional concepts in terms of non-intentional ones (if rationality can ultimately be analysed in non-intentional terms). However, as a form of interpretationism, Child (1994) is right that it is not very ambitious. Derivative interpretationism does not suggest a particularly distinctive or important role for the idea of interpretation to play in the understanding of thought.

The greatest problem for this form of interpretationism is to give the promised analysis: to say something sensible about rationality without talking about thought and interpretation, which then allows us to derive the consequence that all thought is indeed interpretable.

3.2 Analytic interpretationism

The aim of analytic[20] interpretationism is to also to give an analysis of thought in independent terms, but it says that one of the most important of these terms is

[20] This is an unfortunate name for this position, invoking as it does the analytic-synthetic distinction and all its attendant problems, while also recalling both analytic behaviourism and analytic

'interpretability'. According to such an account, having a thought is constituted by being interpretable as having that thought: and one has a thought in virtue of being interpretable as having that thought. The claims above, that thoughts are things which are identifiable by a process of understanding and explaining thinkers and their actions in terms of reasons, and that thinkers are beings susceptible to this kind of understanding, do then give us an analysis of the concepts of thought and thinkers. This is the position that Child (1994) seems to have most prominently in mind when he uses the label 'constitutive interpretationism'.

The question of where rationality fits into this account is a little harder than with the previous account. It seems that there are two possibilities: either the rationality of thinkers might be taken as a consequence of their interpretability, or the notion of rationality might play some role in our account of interpretation. Thus, it is unclear whether, on this view, the Availability Claim is strictly prior to the Rationality Claim or not.[21] As a form of interpretationism, it is clear that analytic interpretationism is highly ambitious: it claims that the notion of interpretation is essential in understanding the nature of thought.

The greatest problem for this form of interpretationism is, as Child points out, that it appears to be circular. In fact, Child suggests *three* circularity objections to this theory: first, that interpretability cannot be explained without reference to the properties people are interpretable as having, i.e. thoughts, which are what we were supposed to be analysing; second, that interpreting someone involves having a propositional attitude about them, and we can ask what it is for an interpreter to have this attitude, thus leading to a vicious regress; and third, that interpretationism tries to explain what it is to be a rational agent in terms of interpretability, but to understand interpretability, we have to use the idea of it being possible to treat the

---

functionalism, even though the theory in question is not necessarily analytic in exactly the sense that either of those two theories are analytic. Nevertheless, I cannot think of another more suitable name.

[21] More on notions of priority shortly.

creature as if it were a rational agent, and so the idea of a rational agent must explain, rather than being dependent on, the idea of interpretability.

One way to react to these threats would be to retreat to derivative interpretationism. However, analytic interpretationism might have the resources to address them. For example, if the notion of rationality is taken as prior to the notion of rational agency, and rationality is part of our account of interpretation, it appears there is more potential for avoiding circularity than if the rationality of thought is merely taken as a consequence of interpretability. Child also suggests that the interpretationist might make use of a distinction between different kinds of priority. He suggests that Peacocke's (1981) distinction between definitional and cognitive priority might be helpful:

> One might, for example, say that the concept of being interpretable as believing that p is *definitionally* prior to the concept of believing that p (i.e. the concept of believing that p can be illuminatingly defined in terms of the concept of being interpretable as believing that p), but that the concept of believing that p is *cognitively* prior to the concept of being interpretable as believing that p (i.e. no one could possess the second concept without possessing the first). (54)

The hope would be that once we recognise the different senses of priority in play, we will see that there is no circle involved in the theory after all. I agree that the notion of priority as it has been used in this discussion so far is sorely in need of clarification, and that there may well be more than one notion in play. It is far less clear to me, however, that these different notions will allow us to break out of all potential circles.

Child also suggests another way, his preferred way, to avoid circularity. He suggests that the interpretationist adopt a 'no-priority' view, and decline to say that either thought or interpretation is prior to the other in any important sense. I will

say that this response amounts to rejecting analytic interpretationism, and moving to either dependence or cartographic interpretationism, depending on whether the project is still conceived as one of analysis.

3.3 Dependence interpretationism

Dependence interpretationism tries to adopt Child's solution to the circularity problem without giving up on the idea that interpretationism provides an analysis of thought. It claims that we can give an analysis of thought in independent terms, but that such an analysis must be an analysis of *both* thought *and* interpretation, because the two cannot be analysed separately.

The argument that Child suggests for constitutive interpretationism could be used to support this view. This is based on Davidson's suggestions in, for example, 'Thought and Talk',[22] and it says that belief only arises in the context of two (or more) creatures interacting with each other in a certain way: a way which involves each interpreting the other. The claim of the dependence interpretationist could then be that for a pattern to involve thought it must involve more than one subject, and the two or more subjects must interact in a particular way. Rationality then enters into the picture when we come to describe the pattern in independent terms, as the dependence interpretationist insists that we can. He may, for example, say that the lives of the creatures must instantiate rational patterns. Again, it should be noted that 'rationality' may be being used in a way that is to some extent non-standard.

Dependence interpretationism is ambitious in terms of what it is trying to achieve (a proper analysis of thought) and as a form of interpretationism (since it gives the understanding of interpretation an important and distinctive role in the understanding of thought). In addition, the dependence theorist won't want to say that one has a thought in virtue of being interpretable as having that thought. Thus, the role of the notion of interpretation in dependence interpretationism is not the

[22] Which can be found in his (1984). I do not mean to claim that Davidson is a dependence interpretationist, however.

same as the role of the notion of interpretation in analytic interpretationism. This picture, then, fits in well with the things Child says about his preferred version of constitutive interpretationism.[23]

While I think that, given Child's definitions and discussion, we probably ought to class this as a form of constitutive interpretationism if we are going to apply the distinction, it is also important to note that in an important respect, dependence interpretationism has more in common with derivative interpretationism than it does with analytic interpretationism. Like derivative interpretationism, it does not count the important interpretationist claims about thought and thinkers being interpretable as giving an analysis of those concepts.

The circularity objections that threatened analytic interpretationism have little bite against this view. It can respond that one cannot give priority to either the notion of thought or that of interpretability; however together they can be specified without adverting to either notion by talking about the independently specifiable pattern involving two or more creatures that constitutes thought and interpretation. You cannot then explain thought without talking about a process of interpretation, but you can explain both thought and interpretation without invoking either.

The important problems for dependence interpretationism are rather presenting a persuasive argument for why thought and interpretation must come as a package, in order to avoid returning to derivative interpretationism,[24] and then producing the analysis of the package, including saying something suitable about rationality which does not invoke the notions of thought or interpretation.

3.4 Cartographic interpretationism

Alternatively, in taking a 'no-priority' view of the relationship between thought and interpretation, one might thereby reject the aim of giving an analysis in independent terms. Instead, one might attempt to map out the relationships between the two

[23] See note 85 on p48 and note 96 on p54 for Child's discussion of the expression 'in virtue of'.
[24] Child, at least, is not convinced by the Davidsonian argument mentioned above.

concepts, thereby illuminating both of them, without performing any sort of conceptual reduction or placing them in a hierarchy. In such an account, rationality then appears as another importantly related concept, on a level with thought and interpretation. I will call this project cartographic interpretationism.

In one sense, this project is less ambitious than the other three, just because it does not aim to give an analysis.[25] On the other hand, as a form of interpretationism the theory is very ambitious: it says that we cannot understand thought without understanding interpretation, and so gives interpretation a central, distinctive and ineliminable role in the understanding of thought.

I suspect that, if we are to employ Child's terminology, this should count as a form of constitutive interpretationism, because of the central role it gives to interpretation. However, it seems inappropriate to say either that one has a thought in virtue of being interpretable as having that thought, or that having a thought is simply constituted by being interpretable as having that thought: both suggest the aim of analysing thought which is explicitly rejected by this theory. In rejecting hierarchy among the concepts of thought, interpretation and rationality, the theory will also refuse to assign priority of a definitional or analytical sort among the Availability and Rationality Claims.

Cartographic interpretationism also suggests a respect in which analytic behaviourism, and the other three types of interpretationism, go beyond the requirements for being in the 'No' camp: it suggests that to be in the 'No' camp, one does not have to offer an analysis of thought and of thinkers, one just needs to say that when explaining what other concepts the concept of thought relates to, ideas about particular kinds of internal organisation of thinkers do not come into the

[25] It might, however, be combined with a claim that such an analysis is not possible and/or that there is no perspective external to the practice of interpretation from which thought can be accurately and completely described. A theory which sought to demonstrate this would have a different, but strong ambition.

picture. This is enough to disagree with the central claim of those in the 'Yes' camp, and so to show that you are a member of the 'No' camp.[26]

Cartographic interpretationism will either reject charges of circularity entirely, or claim that it presents us with a virtuous, because genuinely illuminating, circle of concepts. Its central challenge will be to deliver an account which is, indeed, illuminating.[27]

3.5 Conclusion to section 3

The four-way distinction I have proposed between varieties of interpretationism is necessarily schematic at this stage, but it does make clear how theories which share a commitment to the claims of section 1 may nevertheless differ in structure and aim. I will appeal to these different varieties of interpretationism in the course of my thesis, showing some of the effects that a choice between the different types will have. I do not, however, argue decisively for any one of them. The aim of my thesis is to show that interpretationism, as a broader category of position in the philosophy of mind, is plausible in that:

1. it has the potential to fulfil certain promises of a theory of mind which focuses on the behaviour, rather than on the internal make-up of candidate thinkers.
2. it can fend off certain apparent counterexamples.

As we will see in the next section, there are plenty of challenges to the idea that *any* theory which accepts the Availability and Rationality Claim can achieve this, and I focus on these.

---

[26] Likewise, someone could be in the 'Yes' camp without needing to give a full analysis. As mentioned in chapter 1, section 2, Block may hold such a position.

[27] Naturally, we will now want to ask what it means for such an account to be illuminating. I contend that there are some generally agreed signs of such illumination and suggest that cartographic interpretationism displays these in Appendix 3.

4. Problems for interpretationism

A first problem for interpretationism is proving that it can indeed allow that mental states cause the behaviour that we commonly suppose them to. Exactly how an interpretationist can ensure this is a difficult issue, and one that has received substantial attention in the literature. Donald Davidson tackles the question in his 'Mental Events',[28] which has formed the starting point for many discussions of mental causation, and Child (1994) discusses the issue at length. I, however, do not intend to engage with this literature, but assume that some workable account can be found.

Instead, I focus on what we might call difficult cases for interpretationism, many of which have been presented as counterexamples to the position. These come in two categories.

First, there are cases which challenge the Availability Claim. Some of these suggest that interpretability cannot be necessary for thought. There are extreme cases here, such as the case of the person with a rich mental life who suffers from Total Locked-in Syndrome.[29] However, there are also much more mundane cases where it seems very difficult, if not impossible, for an interpreter to discover another person's thoughts. For example, is there any sense in which all the thoughts occurring within dreams are interpretable, given that they happen during sleep and are then usually forgotten? And could an interpreter always see through a subject's attempts to deceive them? These problems reflect a feature of interpretationism which may seem highly counter-intuitive, namely that it does not seem to make room for our ideas about the privacy of the mental. Other challenges to the

[28] To be found in his (1980).

[29] This is a condition where a person is awake and aware, but all of the voluntary muscles in their body are paralysed. In such cases doctors may believe that their patients are aware only because of evidence from brain-scanning techniques. See for example 'Detecting consciousness in a total locked-in syndrome: an active event-related paradigm' in *Neurocase* 2009 Aug: 15 (4): 271-7.

Availability Claim threaten the idea that interpretability is sufficient for thought. For example, could even the most complex, interpretable behaviour establish that we were dealing with a thinker, if that behaviour was produced in highly non-standard ways, for example by a mechanically simple, though enormous, look-up table?

The second category of apparent counter-examples is supposed to show that the Rationality Claim is implausible. There are various phenomena which suggest that ordinary human thought is not always, or perhaps ever, very rational. Once again, dreams provide an example here: they are well known for often making little sense. Philosophical attention has been also been directed at the results of the Heuristics and Biases Program in cognitive psychology: some have taken these results to show that even in relatively simple, everyday reasoning, most people use unreliable heuristics rather than what one would expect to be familiar logical rules. Other cases include the traditional problems of irrationality, and the thoughts often attributed to people with certain mental health problems, such as those who suffer from delusions. In response to such cases, interpretationism may seem to face an unpalatable choice between saying that these are really instances of rationality, or saying that many 'people' and perhaps even all humans, due to a lack of rationality, fail to be thinkers.

How serious these difficult cases are for the prospects of interpretationism clearly depends on exactly how we are supposed to understand the Availability and Rationality Claims. My plan in the rest of this thesis is to focus on certain of these hard cases, and to use them to develop the crucial notions of interpretability and rationality. I will argue that we can develop these notions so that the Availability and Rationality Claims look plausible, and interpretationism still has the potential to satisfy the three reasons for occupying the 'No' camp that were suggested in chapter 1.

PART II: INTERPRETABILITY

Interpretationism claims that a being has a thought iff it is interpretable as having that thought. The purpose of Part II is to consider threats to this claim, and through doing so to offer the most plausible account of 'interpretability'.

Chapter 3 presents the first challenge to the Availability Claim: the problem of seemingly hidden thoughts which occur during dreams. I use the case of dreams to develop a notion of interpretability according to which a thought is interpretable if either a) there is sufficient evidence concerning the thought in the subject's actual situation and actions, or b) there would be sufficient evidence in at least one suitable counterfactual situation.

In chapter 4, I argue that although it might be plausible to say that each individual thought could be interpretable in this sense, two features of our mental lives (the way our thoughts are influenced by interactions with the world, and our propensity to forget many of our thoughts) indicate that no real interpreter could interpret all the actual thoughts of a subject. I then argue that this need not conflict with another commitment of interpretationism, to the holistic nature of interpretation.

Finally, chapters 5 and 6 address further challenges to the Availability Claim. Chapter 5 uses two difficult cases to make suggestions about what we should count as suitable counterfactual situations, while chapter 6 argues that the interpretationist should accept some rather strange thinkers, and can counteract the intuitions which seem to tell against doing so.

Chapter 3 – The Interpretation of Dreams

1. The problem

We have some thoughts which are at least extremely difficult for an interpreter to discover. Let us call these 'hidden thoughts.' A striking example is provided by those thoughts which occur during dreams.

When the normal person enters REM sleep (the state in which most dreaming occurs), their muscle tone decreases to the extent that, if the same level of muscle atonia is experienced when conscious, it feels like paralysis. One suggestion for why this occurs is that it ensures that we don't act out our dreams. Another important feature of dreams is that although a dreamer's real environment can affect the content of their dream to some extent (for example, the noise of an alarm may be incorporated into a dream, so that the dreamer experiences the noise as coming from a different source), a dreamer's thoughts are far less influenced by their current environment than is the case for waking thoughts. Consequently, two important factors in the interpretation of thought are diminished in the case of dreams: the connection between thought and action, and the connection between thought and environment. As a result, it is probably impossible to interpret most dreams as they are happening. In accordance with our normal sleep cycle, most dreams are then forgotten before or almost as soon as we wake up. Thus, for the majority of dreams, the dreamer cannot even tell an interpreter what the dream involved when they wake up.

Dreams, then, seem to provide a straightforward counterexample to interpretationism's Availability Claim. Every human dreams, and so it seems that

every human has a vast number of thoughts which are not available to any interpreter.

This chapter argues that interpretationism can account for dream thoughts if it adopts the right account of interpretability. This involves considering how the interpretationist should use counterfactuals in his theory. Throughout the chapter, I concentrate on the interpretation of the content of dream thoughts.

Section 2 begins the task by considering three bad solutions to the problem posed by dreams, and thus reviewing some of the desiderata for a good solution. Section 3 considers how the interpretationist might employ counterfactuals in his theory, argues for one particular way of using counterfactuals, and in so doing offers what I take to be the best available understanding of interpretability. Finally, section 4 compares my interpretationist account of dreams with the account given by Norman Malcolm in order to answer questions about when we should say dreams occur, and what relationship we must posit between dreams and dream reports.

2. Some bad solutions

There are a few apparently simple solutions to the problem posed by dreams. First, an interpretationist could claim that dreams are not/do not involve thoughts. They would not then be something that interpretationism needed to account for. This would involve rejecting what Dennett calls the "received view" of dreams:

> Dreams consist of sensations, thoughts, impressions, and so forth, usually composed into coherent narratives, occurring somehow in awareness or consciousness, though in some other sense the dreamer is unconscious during the episode. (1976: 151)

Although rejecting this marks a move away from common sense, current scientific thinking, and philosophical tradition,[30] it is not unprecedented. However, many of the arguments against the received view concentrate on questioning whether dreams are, in any sense, conscious,[31] and when, during the period that the subject is out of communication with an interpreter, they occur.[32] The interpretationist who wishes to get out of needing to account for dreams must also question whether dreams are, or involve, thoughts. He therefore requires a particularly strong rejection of the received view, and one which initially seems to leave us with a difficult task in explaining all the waking beliefs and other thoughts people have about their dreams. If the problem posed by dreams is not limited to dreams, he may also have to deny that a large number of other apparent episodes that we usually count as thoughts are contentful.

Still, such a position is possible.[33] My argument against it is that there is no need for interpretationism to deliver such counterintuitive consequences, and so consists in the positive suggestions of this chapter for how the interpretationist should account for dreams.

A second possibility is a variant on the first. It admits that dreams may involve thoughts, but restricts the scope of interpretationism so that it does not purport to

---

[30] Malcolm (1959) notes that Aristotle, Descartes, Kant, Russell, Moore and Freud have all explicitly endorsed the received view.

[31] See Dennett (1976) for a discussion of whether we are conscious during dreams. The view Dennett suggests in opposition to the view that dreams are conscious experiences does not prevent dreams from possessing content.

[32] The main challenge here arises from the existence of arousal dreams, where it appears that a disturbance simultaneously produces what seems to have been a long dream and awakens the sleeper. Some suggestions for explaining these phenomena can be found in Mullane (1983) and Gregory (1916). For the argument of this chapter, the significance of arousal dreams lies in their ability to raise questions about when dreams really occur, whether we can know when they occur, and what this tells us about the nature of dreams. This is discussed in section 4 below.

[33] I initially believed that it was held by Norman Malcolm in his *Dreaming* (1959). Malcolm's position is more sophisticated than the sketch above implies, however, and I discuss how his proposals relate to my own in section 4.

cover them. However, this severely limits the scope of the theory. For example, it would prevent the interpretationist from being able to claim that his theory tells us general and philosophically important things about mental content as such. Another problem for this solution would be giving a reasonable justification for which thoughts were and were not covered by interpretationism. If the *only* difference is that dreams are not available to interpretation, we may start to wonder whether interpretationism tells us anything philosophically useful about thought: the claim that interpretable thoughts are interpretable is not very interesting.

A third easy solution to the problem of hidden dreams abandons interpretationism as I have characterised it in favour of a theory such as that outlined by Lewis in his 'Radical Interpretation'.[34] This involves saying that the interpreter to whom thoughts must be available is not an ordinary human, but a fictitious mighty knower who knows the subject's movements, the forces acting on them and produced by them, all the light, sound and chemicals absorbed or emitted by the subject, all the material parts of the subject and their movements and so on, the masses and charges of the particles that compose the subject, the magnitudes and directions of any fields or whatever permeates him, the complete history of the subject, the complete history of others like the subject, and the nomological and counterfactual dependences between all of these facts. For someone who possessed all this information, the absence of communication with a dreamer during sleep would present far less of a barrier to interpretation.

However, Lewis's theory involves rejecting interpretationism as I have characterised it. It does not count as a member of the 'No' camp, because it allows that internal brain states might be relevant to a creature's mental state even without having the potential to affect outer behaviour. It also fails to satisfy one of the motivations for being in the 'No' camp: it raises sceptical problems about *our*

[34] Which can be found in his (1983).

knowledge of minds and their nature, because it leaves open the possibility that the mighty knower might have a hugely different view of our mental states from us (about both our own and other people's minds). Therefore, although Lewis's theory admits the importance of the notion of interpretation, adopting his theory amounts to abandoning the purpose of this thesis, as described in Part I.

Let us therefore look for another solution: one which does not deliver hugely counter-intuitive consequences, which retains the potential to say something philosophically interesting about thought, and which fits the brief of Part I.

3. Ways of employing counterfactuals

A fourth suggestion: although most dreams are *actually* forgotten, it seems that they *could* be recalled if the dreamer were to be woken during the dream and immediately questioned. Perhaps, therefore, the content of a dream is determined by what an interpreter *would* learn in such counterfactual situations.

This raises an important question about the use of counterfactuals in interpretationism. Here are three options:

1) The interpretationist could disallow any use of counterfactuals in his theory. Every genuine thought would then need to be not only interpretable, but actually interpreted.
2) The interpretationist could allow counterfactuals concerning the interpreter of his theory, but not concerning the subject, i.e. determining content would involve imagining what an interpreter would have concluded if present, but could not draw upon what the subject would have done in different circumstances.
3) The interpretationist could allow counterfactuals about both subject and interpreter, so that the contents of a subject's thoughts are determined by

what an interpreter would conclude if they knew about what he actually does and/or what he would do in other circumstances.

Although there is evidence[35] that Davidson thought that subjects must actually be interpreted in order to have contentful mental states, this does not mean he was committed to option 1. Davidson's requirement of actual interpretation is connected to his theory about how we acquire the notion of objective truth, and how the meanings of terms in basic occasion sentences are connected to their causes. Even Davidson, therefore, requires only that each person be interpreted at some point, not at every point.[36] It is simply hopelessly implausible to insist that every thought be actually interpreted by a real interpreter.

Option 2 is more plausible than option 1. According to this option, a thought counts as interpretable if there actually is sufficient evidence for the existence and content of that thought. This would be a very reasonable way to understand the word 'interpretable', if told nothing more about it. If interpretationism employs this notion in their Availability Claim, then it must say that the existence of any interpretable thought guarantees the existence of some behaviour which could provide evidence for the interpretation of that thought, although there need not be any interpreter around to observe it. However, option 2 and its attendant notion of interpretability provide no help with the problem of hidden dreams. Even when we counterfactually suppose that there is an interpreter sitting at your bedside, he cannot tell what you are dreaming unless we counterfactually suppose that you wake up.

Option 3 is the one required if we are to adopt the solution to hidden dreams suggested above. It is also the view some commentators attribute to Davidson. For example, Jonathan Ellis says that, for Davidson, 'believes' is a predicate constructed

[35] See for example 'The Second Person' in his (2001).
[36] However, see Byrne (1998) footnote 17 for a suggestion that Davidson is in fact making the stronger claim.

to capture 'aspects of the complicated structure of one's behaviour and dispositions to behaviour.' (2011: 193). Note that adopting option 3 involves departing from the understanding of interpretability outlined in the paragraph above. Instead of requiring actual and sufficient evidence for whatever is being called interpretable, option 3 moves us to a weaker notion of interpretability, which requires only that it must be *possible* for there to be sufficient evidence for whatever is being called interpretable.

Ellis (2011) seems to think that Davidson endorses option 3: he suggests that a subject's behaviour in counterfactual situations is relevant to his actual mental states when he says that, on Davidson's view, 'An interpreter who had complete knowledge of *all of a subject's potential behaviour* and the circumstances under which it occurred would be in the position to know everything a speaker believes.' (2011: 194, italics added)

But Ellis's statement goes beyond the idea that the truth of counterfactuals is relevant to the thoughts possessed by a subject; he is also saying that *knowledge* of these counterfactuals is relevant to and sufficient for *knowledge* of a subject's beliefs. One may then wonder whether Ellis also thinks that knowledge of counterfactuals is also necessary for reliable interpretation. If that is the case, then to know that Seb is thinking that an ice cream would be nice, an interpreter would need to know what Seb is and has actually been doing and seeing, *and* what he would do if there were an ice cream van within his sight which only sold ice cream, *and* what he would do if there was an ice cream van within his sight which sold both ice creams and cold drinks, etc.

Call this view, which asserts an interpreter's need for knowledge of counterfactuals, option 3*. As we will see shortly, this adds something to option 3 as stated above, and is not the only route that the interpretationist can take.

Regardless of what Ellis intended to assert, Byrne (1998) certainly thinks that the interpretationist needs to endorse option 3* in order to explain how actually

hidden thoughts can count as interpretable, and he thinks that interpretationism can be criticised on this basis. He objects that giving our interpreter knowledge of *all* relevant counterfactuals gives her knowledge that a real interpreter cannot have:

> We have to give the Interpreter knowledge of counterfactuals that cannot be known on the basis of how the subject actually behaves. But the ordinary person adopting the intentional stance, or the anthropologist approaching a totally alien tribe, are simply not in possession of enormous quantities of counterfactual knowledge that cannot be gleaned from actual observation of the subject. (1998: 219)

According to Byrne, this dramatically distances the interpreter of interpretationism from real human interpreters (as well as the radical interpreter of Davidson's thought experiments). His point is that a real interpreter *cannot* work out information about lots of different counterfactual situations when such knowledge is not determined by actual observation of the subject, and so option 3* requires an interpreter who possesses information that no real interpreter could have. The interpreter who possesses large amounts of counterfactual knowledge isn't as different from us as Lewis's mighty knower. However, their position is still so far from being attainable that we face a problem with accounting for our everyday knowledge of minds. Moreover, it is unclear how *any* interpreter could get such vast amounts of counterfactual information. On this option, therefore, the interpreter posited by interpretationism is not even a reasonable idealisation of real interpreters. At most, it is a degenerate idealisation, abstracting away from important features of our practices of interpretation.[37] It may then also fail to satisfy the third reason suggested for occupying the 'No' camp in chapter 1, section 3.[38]

[37] Cf. Hooker (1994), which gives an account of useful vs. degenerate idealisations.
[38] I.e. the need to give an account of our concept of thought which fits in with the origins and uses of our psychological concepts.

We might wonder whether the situation here is really as bad as Byrne supposes. After all, we do sometimes know counterfactual information about other people. For example, if we know Seb well, we may be certain that, whenever it is hot, he will be disposed to buy an ice cream if an opportunity presents itself. In that case, on any particular hot day it seems that we can know what Seb would do if he saw an ice cream van, even if there is no such van in sight. Such knowledge is not derived from actual observation of a subject at the moment the thought is to be attributed to them, but it is derived from actual observation of the subject over a long period of time. Perhaps, then, if a real interpreter were to watch a subject for long enough they would have all the information, including counterfactual information, needed to fully interpret the subject?

Unfortunately, this is not the case. For this to provide a reasonable solution to our problem, people and their thoughts would have to be (as a matter of necessity) far more uniform and therefore predictable than they actually are. Some thoughts, such as Seb's desire for ice cream, may be highly predictable. Moreover, particular interpretations of Seb may both commit one to and depend upon certain beliefs about counterfactuals concerning Seb. For example, if I interpret Seb as wanting an ice cream, I may also commit to the belief that if he was offered an ice cream he would accept it (all things being equal), and the fact that Seb has so often wanted ice creams in the past may encourage the belief that he would accept an ice cream now if offered one, which may be part of what (in combination with other evidence) informs my judgement that he does want an ice cream now. Nevertheless, we think that many other thoughts are not predictable from previous behaviour, and that some of them are never actually displayed in subsequent action.

The ordinary interpreter may know what alternative situations would reveal such thoughts, and the fact that these thoughts *could* be so revealed seems to be enough to persuade us that such thoughts do exist. Still, the ordinary interpreter does not know, on the basis of actual observation, that these counterfactual facts obtain. I conclude that what evidence would be available in certain situations, as

well as what evidence is actually available, must be relevant to the determination of some of people's thoughts, including many of their dreams. We cannot defend the position that Byrne (1998) attacks.

We therefore reach the problem that Byrne poses for the interpretationist: the interpretability of a thought cannot depend solely on actual evidence, yet if we allow an interpreter access to 'non-actual evidence,' we distance the position from which thoughts count as interpretable too far from our own position, thus creating a problem for our knowledge of minds.

We can solve this problem by noting the difference between saying that interpretability depends on *counterfactuals about what an interpreter could know*, and saying that interpretation requires *knowledge of counterfactuals*. The first, expressed in my statement of option 3 above, does not imply the second, which Byrne at least thinks the interpretationist will call on in explaining interpretability. The interpretationist, then, should adopt option 3 without adding any claims about an interpreter needing to know counterfactuals which he cannot know on the basis of his actual evidence.

According to this option, a subject is interpretable as having a thought iff either a) there is sufficient evidence concerning the thought in the subject's actual situation and actions, or b) there would be sufficient evidence in at least one suitable counterfactual situation.[39] Since I support and develop the notion of interpretability corresponding to this, I will give it a name: potential behaviour interpretability (PB-interpretability for short).

If we are to take this option, we need to say something about which counterfactual situation is the important one, and whether there is just one

[39] Neither the actual nor counterfactual situations mentioned here should be taken as limited to the place, environment and actions of the subject at the time a thought occurs. As explained in chapter 2 part 2, the sort of interpretation the interpretationist is interested in will use a great deal of information about the subject's environment and behaviour over a period of time. 'Situation' here must be understood broadly enough to allow for this.

counterfactual situation that will do, or whether there might be several possible situations in which an interpreter would have the information they needed to identify a thought. We could claim that there is one situation which constitutes the 'ideal conditions' for interpreting a thought, or that there are often several possible situations which would count as sufficiently good for interpretation. The second suggestion is more plausible, and fits well with Davidson's and Child's assertions that the interpretationist need only say that an interpreter can discover the thoughts of a subject *in favourable conditions*. Either way, we need to say something about what count as 'sufficiently good' or 'ideal' conditions for interpretation, and in doing so we face a threat of circularity. I return to this issue in chapter 5.

Suppose for now that we settle on the idea that a thought is PB-interpretable if either there is sufficient evidence concerning the thought in the subject's actual situation and actions, or there would be sufficient evidence in at least one suitable counterfactual situation. The interpreter who succeeds in interpreting the thought in question on this view appears to be far less removed from a real interpreter. We might say that this interpreter provides a virtuous idealisation of a real interpreter – one which can throw light on and help to explain our interpretive practices, rather than obscuring them. Option 3 therefore seems to provide the best opportunity for resolving the problem interpretationism faces with dreams.

It should be noted that this account of the interpretability of dream thoughts is at least very similar to a suggestion offered in Child (2007), which discusses Wittgenstein's view of dreams.[40] Child suggests that such a view amounts to a form of anti-realism about dream thoughts. In the next section, I develop my interpretationist account of dreams further by comparing it to the account given in Malcolm (1959), which is also inspired by Wittgenstein's discussion of dreams. This allows me to argue that the account of this chapter does provide a genuine

[40] See (2007: 255). Child does not attribute the view to Wittgenstein, but offers it as an account which he did not consider.

alternative to simply denying that dreams involve contentful mental states (the first bad solution in section 2 above), to elucidate the relationship that interpretationism should posit between dreams and dream reports, and to comment on the issue of when dreams occur. It will also indicate possible reasons for not taking my account of dream thoughts to be an anti-realist one. However, I will not argue directly for the idea that my view does not involve at least a moderate form of anti-realism about dream thoughts.

4. Malcolm on dreams

Considerations about dreams and how we know about them appear in Wittgenstein's *Philosophical Investigations*, and have been discussed by some of his followers. These discussions have often proceeded from an attempt to understand how we apply the concept of dreaming to conclusions about how we should describe the phenomena, an approach that should be amenable to many interpretationists given the third reason for being interested in interpretationism as suggested in chapter 1.

In his (1959) book *Dreaming*, Malcolm argues for a strong connection between dreams and dream reports; he says that the latter are the sole criterion of the occurrence of the former. He then uses this, among other commitments, to argue for his view of dreams. This is summed up by Schroeder (1997) in 3 controversial theses:

> (1) The temporal location of dreams as taking place in one's sleep is not an empirical fact, but determined by grammar. (D, p.50)
> (2) This grammatical determination does not allow dreams a precise date in physical time (beyond saying that they occur during one's sleep). (D, p.70)
> (3) Dreams do not consist of mental occurrences such as thoughts, sensations, emotions, images, hallucinations &c. (D, p.51f.) (Schroeder, 1997: 15)

So, Malcolm's idea about the connection between dreams and dream reports bears some resemblance to the claims of this chapter, and moreover his arguments stem from a methodology which at least resembles that of my interpretationist. We must therefore ask: is my interpretationist committed to any of these theses? And if not, what should he say about Malcolm's arguments for these claims?

My interpretationist does not want to say that dreams are not mental occurrences: such a claim would amount to the first easy but unsatisfactory solution to the problem of dreams discussed in section 2. However, although some have taken this as Malcolm's claim, he does not state his own position quite so baldly. For example:

> I was inclined at one time to think of this result[41] as amounting to a proof that dreaming is not a mental activity or a mental phenomenon or a conscious experience. But now I reject that inclination. For one thing, the phrases 'mental activity', 'mental phenomenon', 'conscious experience' are so vague that I should not have known what I was asserting. (52)

Another problem with the bald claim is that it could be made false by simple stipulation of the definition of a mental occurrence. Instead, Malcolm therefore claims:

> if anyone holds that dreams are identical with, or composed of, thoughts, impressions, feelings, images, and so on (here one may supply whatever mental nouns one likes, except 'dreams'), occurring in sleep, then his view is false. (52)

[41] This refers to the result of the arguments Malcolm gives in chapters 9, 10 and 11 of *Dreaming*. What I take to be the most important part of Malcolm's argument is reproduced below.

Given this, it seems that Malcolm *may* want to draw a distinction between dreams and waking thought without denying that dreams have content, which has been the central issue of this chapter. I think it is unclear exactly how Malcolm's third claim is to be cashed out, but he does certainly draw a distinction between dreams and waking thoughts which has not been drawn by my interpretationist.

Malcolm supports this distinction between dreams and waking thoughts using an argument which is also supposed to establish claims (1) and (2). Since dream reports are the only criterion for the occurrence of dreams, there is supposed to be a logical link between dreams and reports which, according to Malcolm, prevents us from being able to characterise dreams independently from the reports at all.[42] But this then means that we can't think of the dream as unknowably occurring at one time rather than another. Since we cannot tell on the basis of dream reports the difference between the situation where dreams occur during sleep and then are then remembered on awakening, and the situation where dreams only occur during awakening,[43] dreams are not supposed admit of this difference. It is a feature of the way we talk about dreams that we say they occur during sleep – and as such it is part of our grammar. But since we cannot say when they occur, they do not actually have a precise date in 'physical' time, and nor is the claim that they occur during sleep an empirical claim. Thus we arrive at the first two claims above.

This argument moves from our epistemological situation to what appear to be metaphysical claims about the nature of dreams. This sort of a move is sometimes legitimate according to interpretationism as characterised in chapter 2. However, we must question both if Malcolm gets the epistemological situation right, and if this particular transition is acceptable.

One response to this argument would be to claim that it does not put dreams on a different epistemic footing from many waking thoughts. Schroeder (1997)

---

[42] Some have taken this to mean that Malcolm denies the existence of dreams, or actually identifies them with dream reports. Again, Malcolm would not accept such a bald statement of his position.
[43] The latter is Gregory (1916)'s explanation of the possibility of arousal dreams, which were mentioned in footnote 32.

offers this reply, and uses as examples thoughts that occurred to a subject in the past when he was unobserved, but which we nevertheless accept that he is able to report to us.

The difference that Malcolm would point to in reply is that although in the case of waking thoughts, a report might be the only criterion of the existence of a thought's existence, still that report at least *can* occur at the same time as the thought. In the case of dreams not only do we have a single criterion, but it also can't occur simultaneously with the dream. This is supposed to be significant because of the following principle endorsed by Malcolm:

> I am assuming it to be an a priori truth that whenever a thought occurs to a person or he experiences a feeling he could, at that time, give expression to the thought or feeling. I mean that this is always logically possible (1957: 207)

This then gives us the distinction that claim (3) suggests we should recognise between waking thoughts and dreams.

We could reply to this as Schroeder does, by saying that 'it is not clear what importance verifiability *in principle* may have in a given case where, as it happens, there is *no* possibility of carrying out any checks.' (1997: 29) This, however, would be a very dangerous reply for my interpretationist to make. The notion of interpretability developed in this chapter depends on the idea that it matters if an actually hidden thought can be interpreted in principle, where this means in some suitable counterfactual situation. The notion of possibility that is important to my interpretationist cannot be the one that Schroeder applies, which allows that simply leaving a person alone, or gagging him, removes the relevant sort of possibility of the subject giving a report on his mental state.[44] In making Schroeder's reply my

[44] The sort of possibility the interpretationist requires is discussed further in chapter 5.

interpretationist would run the risk of allowing the problem with dreaming to infect the interpretation of many other thoughts. This would not be to present a 'companions in guilt' style response to dreams, but to admit that interpretationism just doesn't give a good account of a great number of thoughts.

My interpretationist should accept that, if waking thoughts can in principle be expressed at the time they happen while dreams cannot, then this is a genuine difference between the two. However, I see no reason why they need to accept Malcolm's principle above, and so accept that this difference *makes* the difference Malcolm claims it does. Malcolm's principle requires argument, otherwise we can accuse him of simply begging the question against the similarity of dream thoughts and waking thoughts. Malcolm does not appear to give any such argument, and there are good reasons to doubt the principle. For example, anyone who for any reason thinks that not all thoughts can be linguistically expressed by their subjects has reason to doubt the principle.[45]

Perhaps Malcolm would reply that the principle holds precisely because, without it, it might be impossible to know exactly when many thoughts occur. Whether this is true, and whether it matters, will be discussed shortly. First, I challenge the idea that dream reports are the only criteria for the occurrence of dreams.

Why should we think that we have only one criterion for the application of the concept of dreaming, when people sometimes do all sorts of things in their sleep? They may adopt happy, sad or fearful facial expressions; they may start slightly as if surprised; they may laugh or moan. People sometimes also talk in their sleep, and may even perform 'actions' such as walking. We take such incidents as evidence that a subject is dreaming, so shouldn't they count amongst the criteria for the

[45] For example, because they believe non-linguistic creatures can have thoughts. This, of course, raises the question of what the interpretationist should say about inexpressible thoughts within dreams, or the dreams of non-linguistic creatures. I shelve this issue until Part III, chapter 11.

occurrence of dreaming? Malcolm's reply to such considerations is to deny that in these cases the subject is properly asleep, and so to deny that whatever thoughts such phenomena are evidence for count as dreams.

To make this reply, Malcolm employs very strict criteria for what counts as really being asleep. He takes the criteria we teach children when we teach them how to apply the word 'asleep' – closed eyes, lack of movement, obliviousness to mild external stimuli, etc. – and demands that *all* of them be satisfied without qualification for a person to count as asleep rather than merely in a state resembling sleep. This, however, is not how we commonly use the term 'asleep'. Rather, we seem to require only that a sufficient number of the criteria be satisfied, so that if for example a person has their eyes closed and doesn't respond when spoken to, we count them as asleep even if they are smiling and mumbling.[46]

At this point, Malcolm might say that even if we admit that sleepers do not have to fulfil all of our criteria, expression of emotion, genuine speech and other actions have to involve an element of intention or self-awareness which must be lacking in the case of any apparent expressions or actions during sleep. Sleep talking, then, could happen, but it would not be speech; only noise that sounded like speech. To reply to this we could once again raise doubts about whether this got our use of the word 'sleep' right (or our use of the words 'expression' and perhaps even 'action'). However, even if we were to agree with Malcolm on this point, we could still say that these merely apparent actions and expressions could be counted as informative about whether dreaming was occurring, and what it involved.

It seems reasonable to say that there are quite often events, other than dream reports, which count as criteria for the occurrence of dreams. In this respect, dreams are not in fact different in kind from waking thoughts: it is just a matter of the frequency and degree of supporting evidence for such reports.

[46] This seems to fit far better with Wittgenstein's reflections on family resemblances than does Malcolm's approach.

'But,' Malcolm might say, 'there's no way to tell the difference between the case where you *really* speak in your sleep, and the case where you happen to make an *apparently* sensible series of noises which do not correspond to any thoughts. But then there can't be a difference, and such phenomena can't provide evidence for the claim that there are genuine thoughts during sleep.'

In response to this, we could point to the fact that people's behaviour during sleep may agree or disagree with dream reports they give after waking or being woken, and that this provides something of a check on the reliability of both. Suppose a person was tossing and turning in their sleep, and then screamed aloud, whereupon you shook them and woke them. If the person said they dreamt they were being chased and then attacked by a giant spider, this would confirm the idea that they had been having certain thoughts while sleeping. If they complained that you had woken them, saying (with perfect sincerity) that they had been having a lovely dream about sunbathing, then you might conclude that either their behaviour during sleep, or their subsequent report, did not reflect their thoughts during sleep.

In the case where behaviour and report don't agree, of course, it is at least difficult to tell which we should take as indicating the thoughts of the dreamer. And if we cannot determine which we should pay more attention to, Malcolm will say there can't be a difference between the case where the report is inaccurate and the case where the behaviour is misleading. The interpretationist has at least two options at this point. They could agree with Malcolm, *just* for the rare cases where in the actual world or in some suitable counterfactual situation a subject exhibits sleep behaviour and gives dream reports that contradict each other, and there is no way to choose which better reflects their thoughts using principles of interpretation that call on considerations such as rationality. They would then presumably say that the dream thoughts in such cases were massively indeterminate in attitude, content, or both. Alternatively, they could say that in such cases, we give or should give one of the criteria priority. I will not decide between these options here.

Finally, we come to the issue of when dreams occur. One thing we can now say is that we do have good evidence that dreams sometimes occur during what we would usually call sleep and at specific times: namely those incidents in which behaviour during sleep and dream reports on awakening confirm each other. But what should we say about the times when there is no behaviour during sleep, i.e. those occasions when the person just lies there? Perhaps there is a sense in which it is possible that the subject could exhibit behaviour which would confirm their potential dream reports, even though they don't, just as it's possible that a person could have given a dream report if awoken, even though they didn't.

We do not have to rely on this answer, however, in order to say that these dreams do have a precise date in time. This is because, whereas Malcolm seemed to posit a logical link between dreams and actual dream reports, my interpretationist suggests a link between dreams and the disposition to give a dream report if woken. As explained in chapter 3, the interpretationist does not need to identify dreams with these dispositions. Nevertheless, the link posited means that they can call on dispositions to show that a dream could occur at a precise moment in time, just as long as we think that dispositions come to exist at a particular time. The point at which interpretation becomes a possibility then determines the temporal location of the dream. Of course, this does not mean that in any particular case an interpreter can determine when a dream occurred: whenever they wake the subject and ask about their dreams, they don't have a way of knowing the age of the disposition they activate. Thus, everyday interpretation may not be able to tell us when each dream occurs.

The interpretationist account of dreams that I offer thus differs in several ways from the account given by Malcolm. My interpretationist rejects all three of Malcolm's controversial claims, and also suggests a slightly different relationship between dreams and dream reports, as detailed above. Given these differences, the

interpretationist account is not open to all the problems Malcolm's account faces, and certainly should not be taken as simply denying that dream thoughts exist.

5. Conclusion

The apparent 'hiddenness' of dreams prompts us to consider what the interpretationist can plausibly mean when he claims that thought is interpretable. This chapter has argued that the most appropriate notion is that of PB-interpretability: it must be possible for there to be sufficient evidence for the existence and content of whatever is being called interpretable.

In the next chapter, I argue that adopting this notion of interpretability has the consequence that no real interpreter could possess the information needed to interpret all the actual thoughts of a subject, either over an extended period of time or perhaps even within a particular moment. I then argue that this does not present a problem for the interpretationist.

Chapter 4 – The Holistic Nature of Interpretation

In chapter 3 I considered a number of ways of understanding the 'interpretability' of thought, and settled on an account which I called PB-interpretability. Saying that all thoughts are PB-interpretable amounts to saying that for each thought, either there is sufficient evidence concerning the thought in the subject's actual situation, or there would be sufficient evidence in at least one suitable counterfactual situation.

In section 1 below, I argue that while it might be plausible that each thought is PB-interpretable, it is not plausible that all could be interpreted in the same possible world by a real interpreter or a suitable idealisation thereof. Section 2 then draws out the significance of this claim by showing how it connects with a famous claim from Dennett: '*all there is* to really and truly believing that p (for any proposition p) is being an intentional system for which p occurs as a belief in the best (most predictive) interpretation.' (1987: 29) I argue that we should not endorse Dennett's claim as it stands, and thus illustrate one potential difference between Dennett's position and the one developed in this thesis. The conclusion of section 2 is that there is, prima facie, a tension between accepting the Availability Claim (understood as asserting the necessity and sufficiency of PB-interpretability for thought) and a holistic view of interpretation. In section 3 I first explain why a genuine conflict here would be a problem for the interpretationist, and then argue that the apparent tension can be dissipated.

1. The directability of thought

The Availability Claim, if it uses the notion of PB-interpretability, entails that each thought that a subject has must be PB-interpretable. So, for any thought that a subject has, a real or suitably idealised interpreter could identify that thought if she used the right method and was lucky enough to do so in favourable or suitable conditions. But this doesn't guarantee that a real or suitably idealised interpreter could have experienced favourable circumstances and thereby interpreted all the actual thoughts of a particular subject. And as it happens, this final claim is implausible, because of a feature of thought which we might call 'directability': its course, and our memories about its course, can be easily influenced by interaction with an interpreter.

This is particularly clear in the case of dreams. Favourable circumstances for interpreting dreams appear to involve the interpreter knowing the language of the subject, and waking the subject and asking them what they have just dreamed at regular intervals. But this means that the subject will have at least slightly different dreams than he would have done had he been allowed to sleep uninterrupted. It is also likely to change the waking thoughts of the subject: the subject will have more waking thoughts at different times, due to being woken, and some of them will probably be caused by the waking. Moreover, consistently waking a subject when they start to dream has deleterious effects on mental and physical well-being, suggesting that such questioning would drastically alter the majority of the subject's waking thoughts. Although this phenomenon is dramatically illustrated by the case of interpreting dreams, it is in fact widespread. For example, I cannot remember every thought I have had during a long bike ride once I reach home, and so if an interpreter asks me what I was thinking about, I cannot tell them everything. If the interpreter rode along beside me, I could have answered questions about what I was thinking as we went along. That, however, would have significantly altered the

course that my thoughts took. In general, interactions with an interpreter, and other changes in the circumstances of the subject, affect the thoughts that the subject has.

So far, the issue I have raised amounts to the claim that although each thought may be interpretable at or just after the time it occurs, many thoughts arise only because the prior thoughts of a subject were not interpreted. The actual thoughts of a subject, taken together, are not jointly interpretable. I have allowed that an interpreter might have been able to interpret everything a subject thought given the right conditions, but pointed out that this would have involved interpreting different thoughts. We therefore have a problem with saying that *all* the thoughts in a set that extends across time are available to an interpreter. However, we can also produce problems of synchronic availability. These suggest that it may be false even that a real interpreter could, if they were extremely lucky and always observed their subject in favourable conditions, succeed in interpreting some complete set of thoughts of that subject.

Suppose that subject x is dreaming that he is in Vancouver Aquarium when interpreter y wakes him up to ask what he is dreaming. Like many dreams, x's dream is highly detailed and complex. At the moment he awoke, it included the following details: that he was standing in front of a large window with fish on the other side; that there was a particularly large, grey, and ugly fish swimming to the left of him, which he considered to be rather threatening, and five small colourful fish to the right, which he thought very pretty; that he was with his niece, but she was behind him, looking through another window; that some young boys were shouting in the background; and finally that he was worrying whether it was ethical to visit this aquarium, given the way it acquired its beluga whales. When y wakes x up and asks him what he was dreaming, x replies that he dreamt he was in an aquarium. Y then presses x for more details: was it any particular aquarium? Was he with anyone? What could he see? What was he feeling and thinking? Depending on the order in which y asks these questions, x may remember different features of his dream, since during the time it takes to relate some of these details, it

seems likely that others will be forgotten. For example, x might be able to recall the colours of the different fish if he is asked about these first. However, he may forget them in the time it takes to explain why he was worried about being at the aquarium. On the other hand, if x spends the first few minutes after waking listing the colours of the fish, he may forget that there were boys shouting in the background.

This problem of synchronic availability is once again easy to illustrate using the case of dreams, because of the memory loss associated with dreaming. However, it again seems likely that it is more widespread. If we allow that people can have several thoughts at one time, and that they can forget these thoughts, or might not be aware of some of them, then it seems as if with respect to waking thoughts, an interpreter might be able to discover any particular thought, but not all the thoughts that a person has at one time.

Let us return to the original question of chapter 4: how can the interpretationist account for contentful, hidden dream thoughts? The suggestion was that dream content is determined by what an interpreter would learn in counterfactual situations. There was a constraint on this solution: we wanted to avoid demanding that our interpreter have information that a real interpreter couldn't possibly possess. I therefore suggested that there must be enough evidence in at least one counterfactual situation for an interpreter to identify the thought. Now, I have suggested that because thought is directable, no real interpreter could possess the information needed to interpret all the actual thoughts of a subject, either over an extended period of time or perhaps even within a particular moment.

We might question this result, for example by suggesting that it relies too heavily on the contingent fact that we often forget our past thoughts. We might suggest that a real interpreter could possess the necessary information, and could interpret all the actual thoughts of a subject, if that subject had a better memory. We might then suggest that the possession of such a memory on the part of the subject is a component of there being 'favourable conditions' for interpretation. According to

this proposal, what we need to do is to add an additional counterfactual supposition about the subject of interpretation. However, in chapter 5, where I discuss what we should count as suitable counterfactual situations, I argue that we must not include massive changes to the mental capabilities of the subject of interpretation. I claim that at least some failures in memory are an important aspect of our finitude, and as such must be taken into consideration by the interpretationist.

Rather than questioning the results of this section, I will argue that they do not present a serious problem for interpretationism. First, however, we must consider why they are significant at all.

2. Dennett's famous claim

In *The Intentional Stance*, Dennett makes what he calls a 'perverse' claim, namely that '*all there is* to really and truly believing that p (for any proposition p) is being an intentional system for which p occurs as a belief in the best (most predictive) interpretation.' (1987: 29) It is because of the words in italics that he thinks this claim is 'apparently shallow and instrumentalistic' (ibid.), although of course he thinks that it turns out on closer inspection to provide a robust theory.

To this part of the claim, my interpretationist may agree: they too might think that *all there is* to having various thoughts (from a philosophical point of view) is being interpretable as having them. However, Dennett's claim might also be taken to imply that there is some *one* interpretation which details absolutely everything that a subject thinks. And since for Dennett the intentional stance is one that we normal humans adopt in interacting with each other, and which is justified by its usefulness in generating predictions that we can use, this suggests that Dennett may

disagree with the claim that not all thoughts can be jointly available to a real interpreter.[47]

My interpretationist may admit the existence of a single 'interpretation' which gives an account of all the thoughts that a subject has (call this the complete theory), but should insist that this won't be available to any real interpreter. Rather, it will be made up of the combination of the many interpretations which were individually available to a real interpreter.

However, if the complete theory is not available to real interpreters, then those interpreters will not be able to call on considerations about how their interpretations fit into the complete theory in deciding which are best.

To illustrate this consequence, let us return to the difference between options 3 and 3* for using counterfactuals in explaining interpretability, as presented in the previous chapter. According to option 3, to interpret a subject an interpreter requires information from just one situation, and it is possible that he could have been in this situation, although he might actually be in a situation less favourable to interpretation. This was the option I argued for in the previous chapter. In contrast, according to option 3*, interpretation of a subject requires a large amount of counterfactual information, including several counterfactuals about the subject's actions at a given moment. If we had chosen option 3*, we would not be having the current disagreement with Dennett. If the relevant interpreter of a subject was one who had access to all counterfactual information about that subject, then we would have none of the reasons presented in the previous section for supposing that thoughts may fail to be jointly interpretable: since the notion of interpretability in play would not be linked to what a real interpreter could find out, considerations about how many and which thoughts a real interpreter could find out about together

[47] The quote does not provide decisive grounds for this interpretation of Dennett. However, I am not primarily concerned with interpreting Dennett, and regardless of Dennett's actual position it is useful to contrast this potential interpretation with my own view.

would not be relevant. This would then mean that the relevant interpreter could work out which individual thoughts to attribute by seeing how they fitted into the best complete theory of the subject's mental life.

So, if we adopt the notion of interpretability I have suggested, we have reason to believe that not all actual thoughts of a subject are available for interpretation together, and that the *complete* theory of a subject's thoughts cannot guide an interpreter in deciding which individual thoughts should be attributed to a subject. This sounds like a rejection of some kind of holism. In section 3, I address this issue. I first distinguish between some different types of holism, then explain why a retreat to atomism would be disastrous for interpretationism, and finally argue that a commitment to the Availability Claim and to using the notion of PB-interpretability is compatible with the moderate form of holism about interpretation that interpretationism requires.

3. Holism

There are various different things that philosophers might refer to using the term 'holism'. One common distinction is drawn between holism concerning the mind, and holism about language. The former of these is our main concern here. It too can be split into further types, three of which are given below.

Holism concerning the mind might refer to:

A. The view that the concepts we apply to mental states cannot be described or understood separately. According to this view, there is at least one *area of discourse* in which meaning is holistic. However, without further argument it does not necessarily follow that there is anything holistic about the nature of mental states themselves.

B. The view that thoughts cannot be determined one at a time, nor actions given intentional descriptions in isolation, but rather they have to be worked out together.

C. The view that the content and/or attitude of any mental state depends upon the set of which it is a part.

Most detailed discussions of holism seem to focus on type C above,[48] and this is also a brand of holism which Davidson seems to endorse. However, it is primarily type B which should concern us here. I do not discuss type C holism further in this chapter.

Type B holism is important because it is the form of holism challenged by the ideas presented so far in this chapter. Yet it may also seem to be assumed by the process of interpretation which, in chapter 2, I said that interpretationism was interested in. The method placed centre stage by interpretationism is one of working out what an individual thinks by working out how their environment, actions and thoughts must fit together and influence each other. The topic of how interpretationism should claim thoughts fit together is set aside until the discussion of rationality in Part III. However the mere fact that interpretation proceeds on the basis of an assumption about how they fit together is enough to show that interpretation cannot be atomistic.

The emphasis on the importance of a holistic process of interpretation also marked one of the important differences between interpretationism and analytic behaviourism. If interpretation did not proceed holistically, then surely there would be something to say about the behaviour associated with particular thoughts, as analytical behaviourism supposed. Yet it seems there is nothing useful to say here, as shown by many critics of behaviourism. We do not want interpretationism to face these criticisms as well.

---

[48] See Heal (1994) and Jackman (1999) for useful discussions of the different strengths of type C holism available, and thoughts about which seems most plausible.

The apparent conflict is easy to resolve without either turning to a new method of interpretation or collapsing back into something like analytic behaviourism. Our interpretationist should claim both that individual thoughts are available to interpretation without all the surrounding thoughts being available to interpretation, and that we have to determine thoughts in combination rather than separately. But although these two commitments might initially appear to point in different directions, they are not incompatible as long we do not say that *all* thoughts must be determined together. My interpretationist, then, should allow that one does not need to interpret absolutely all the thoughts of a subject in order to interpret one of them, while maintaining that one will often (perhaps always) need to interpret several at a time.

This is the important claim the interpretationist must make in order to avoid a clash between the Availability Claim and their view of the nature of interpretation: you don't need to interpret every single thought of a subject to be able to interpret one of them, but you do need to interpret more than one at a time. This claim seems very plausible, and I presume that many philosophers with different background theories would accept it. My interpretationist, then, is committed here to a reasonably uncontroversial claim. However, his position becomes both more complicated and rather more idiosyncratic when we work out the details of which thoughts do and do not need to be interpreted together according to his view. This provides us with the more interesting puzzle raised by considerations about the directability of thought.

Our consideration of which thoughts need to be interpreted together might follow the pattern of discussions about holism and molecularism or localism with regard to language.[49] In that case, one obvious suggestion would be that we need to interpret all the thoughts concerning a certain domain together. So for example, we

[49] See for example Peacocke (1997).

might suggest that all thoughts about maths have to be interpreted together, and all thoughts about the mental have to be interpreted together, but that to interpret all thoughts about maths you do not need to interpret all thoughts about the mental. There would be some obvious problems with such a proposal, for example the difficulty in delineating domains. Perhaps these problems could be solved, perhaps not; but in any case, this is not what my interpretationist ought to say. My interpretationist needs to rule out the possibility of thoughts which are not co-interpretable needing to be interpreted together, and given the reasons for thinking some thoughts might not be co-interpretable (as presented in section 1) these might well sometimes be thoughts on the same topics. Ruling out this possibility is a project which I do not think other philosophers have addressed, since it is a problem quite specific to my brand of interpretationism.

So, might thoughts which are not co-interpretable need to be interpreted together? Or in other words, might certain mental states fail to be jointly accessible with those mental states whose interpretation they affect? The form of such a counterexample would have to be as follows: It would be a case where, to determine thought a, y needs to know that it occurs in the given situation along with thought b. However, it must also be impossible for y to determine the contents of both thought a and thought b, because if he discovers the content of b, he prevents himself from being able to find out about a and vice versa. Y therefore won't be able to determine the content of thought a. If this situation is possible, then not every thought must be PB-interpretable.

Hopefully this gives an idea of what would constitute a counterexample to the interpretationist's view. It is worth also bringing out some of the complexities which would make it difficult to present the interpretationist with such a counterexample. First, it should be noted that, given the considerations from section 1, the interpretationist does not need to say that any two thoughts are necessarily jointly uninterpretable. Take the example of the dream about the aquarium: any two of the details from the dream may be recallable together, even if all are not recallable

together. Second, cases probably usually involve groups of inaccessible thoughts affecting the correct interpretation of other groups, rather than single thoughts making a great difference to each other. Scenarios are also particularly difficult to construct because we are not supposed to look at actual situations where an interpreter is in fact unable to identify a thought because another remains hidden. Rather, we need a situation with a certain counterfactual structure such that the above *has* to obtain. A counterexample must show that there are some thoughts which can't be interpreted even in the most favourable conditions possible.[50]

I have not managed to think of any situation of this structure which would arise because of the directability of thought and/or the frailty of memory commented on in section 1. I therefore cannot present such a counterexample as evidence against my interpretationist's claim that every thought is PB-interpretable because each thought is co-interpretable with enough other thoughts to establish its existence and content.

What is the status of this claim? For example, should we see it as some sort of hopeful statement of faith on the part of the interpretationist? Or perhaps as simply part of his definition of thought? The former description might seem more accurate to the extent that the interpretationist is depending on the fact that he cannot think of a counterexample, and hopes that no one else will be able to either. The latter would seem more appropriate if the interpretationist would be inclined to mistrust the importance of any intuitions behind supposed counterexamples, and claims that he has deduced that every thought is co-interpretable with enough other thoughts to establish its existence and content as a consequence of the nature of thought.

The best position for an interpretationist, however, would be to be able to produce an argument in favour of the claim. I think that such an argument might be available, and that it might depend on the vast number of thoughts that each person has and is capable of having, and the correspondingly small difference that any one

[50] The complicated nature of the counterexamples required makes it uncertain whether such a counterexample could have much intuitive force against my interpretationist.

thought must surely make to the interpretation of another in all favourable circumstances. This argument could not be pursued, however, without a discussion of how thoughts need to be related to each other, such as that offered in Part III. It will not be developed further in this thesis.

4. Conclusion

The Availability Claim (understood using the notion of PB-interpretability) and the holistic nature of interpretation may appear incompatible at a superficial glance. However, the tension disappears if we allow the quite uncontentious claim that you do not need to interpret every single thought of a subject to be able to interpret one of them, but you do need to interpret more than one at a time. In combination with the interpretationist's theory, this leads to a more unusual claim, namely that that every thought is PB-interpretable because each thought is co-interpretable with enough other thoughts to establish its existence and content. Although far fewer people are likely to endorse this claim, I have not found a counterexample to refute it. I therefore provisionally conclude that my interpretationist can endorse the Availability Claim and a moderate holism about the nature of interpretation.

Chapter 5 – Possibility, Deception and Paralysis

In chapter 2 I characterised interpretationism as being committed to an Availability Claim: the claim that interpretability is both necessary and sufficient for thought. In chapter 3, I argued for a particular understanding of interpretability: PB-interpretability. When we plug PB-interpretability into the Availability Claim, we get the claim that for every thought, either it can be interpreted given the information available to a real interpreter in the actual world, or it could be interpreted given the information that would be available to a real interpreter in at least one suitable counterfactual situation. In this chapter, I consider these 'suitable counterfactual situations', and the plausibility of an Availability Claim which refers to them, in more detail. In particular, I consider the related issues of how the interpretationist can pick out the suitable situations he needs for his theory, and what these situations will be like.

Section 1 explains the nature of this task, and offers some conditions on an account of suitable situations. Section 2 then considers a general problem that the interpretationist might be thought to have in referring to these suitable counterfactual situations as part of his theory, and shows how it can be overcome. The third section considers a particular kind of case – deception – where it may be thought difficult to explain what goes wrong with attempted interpretation in the actual world. Finally, the fourth section considers a case, involving a paralysed thinker, where it might be thought that there are no suitable counterfactual situations for the interpretationist to call on. Both apparent counterexamples are used to characterise suitable situations further.

It will emerge through the discussion of these problems that different kinds of interpretationism have different requirements for this part of their theory. I will argue that, given the considerations of this chapter, cartographic interpretationism is left in the strongest position. I will not give necessary and sufficient conditions on suitable counterfactual situations (in section 1 I suggest this is not needed for the theory to succeed). More significantly, I will not prove conclusively that the proposed counterexamples to interpretationism can be disarmed. However, I will aim to make it *plausible* that, in the case of thoughts hidden in this world, there is always a suitable counterfactual situation where they are not hidden.

1. The task

For those comfortable with possible worlds, a suitable counterfactual situation may be thought of as a possible world (or, alternatively, a part of a possible world, since not everything that happens in the world will be relevant to interpretation of a given thought). I will utilize the language of possible worlds here for ease of expression.

The interpretationist asserts that when a thought is hidden in this world, there is a possible world in which the thought is not hidden. Moreover, he refers to this possible world as part of his account of that thought. For example, he might say 'thought m counts as interpretable because there is a possible world, w, in which there is enough evidence to allow an interpreter to identify it.' This possible world must then be such as to make it clear why the thought should count as interpretable in the actual world. Only such possible worlds will provide *suitable* counterfactual situations.

In this section I consider, first, conditions on how the interpretationist should refer to these counterfactual situations as part of his theory, and why the analytic interpretationist faces a particularly difficult task here; and second, what we need to say about what the suitable counterfactual situations must be like. This will set up the task of addressing these issues more fully through the rest of the chapter.

First, consider the interpretationist's reference to suitable counterfactual situations as part of his theory. Since the interpretationist refers to these worlds as part of his account, he needs to be able to do so in a way that doesn't undermine the account, either by making it circular or by making it trivial. This means that he cannot simply say that the suitable counterfactual situations are the ones where the thought is available to interpretation, i.e. where there is enough evidence to interpret the thought. If he did this, he would be proposing that thoughts are interpretable if they can be interpreted in a suitable situation, and that a suitable situation is one where they can be interpreted. For the analytic interpretationist, who hopes to give an analysis of thought in terms of interpretability, this leads to vicious circularity. However, even if we do not want to analyse thought in terms of interpretability, but rather want to pursue one of the other forms of interpretationism, this suggestion is unacceptable: it gives us a very small and unilluminating circle which tells us nothing about what the suitable situations are like, what interpretation depends on, how close the notion of interpretability in question is to an everyday conception of what counts as interpretable, or how to evaluate the truth of claims that something is or is not interpretable. Anyone who wants to say that the Availability Claim is interesting needs to give a more substantive account of what counts as a suitable counterfactual situation.

All interpretationists, then, face some work in picking out suitable situations in an illuminating way. However, the different sorts of interpretationist may still adopt different strategies. For the analytic interpretationist, suitable situations are part of the analysis of interpretability, which is part of the analysis of thought. Therefore, the analytic interpretationist cannot refer to thought or particular thoughts in picking out suitable situations without circularity.

On the other hand, referring to a mental state in explaining interpretability is far less obviously problematic for the interpretationist who wants to say that a subject has a thought iff they are interpretable as having that thought, but who then

refrains from giving an analysis according to which being interpretable as having a thought is precisely what having that thought consists in.

According to derivative and dependence interpretationists, there is another analysis to give of thought, and this will allow us to explain what it is to have a thought, and will have the consequence that every thought is interpretable. Since the notion of interpretability is not required for the analysis, perhaps we can explain what interpretability amounts to using certain mental states. Indeed, the dependence interpretationist, who thinks that the notions of thought and interpretability have to be understood together, would expect nothing less.

According to cartographic interpretationists, there is no analysis of thought to be given using independent terms, but we can illuminate the nature of thought by showing how it relates to other concepts, in particular that of interpretation. These theorists may accept that there is a circularity involved in the account, but they must argue that it is not thereby problematic (see Appendix 3). They can refer to thoughts when describing suitable situations in general or when describing the situation where an individual thought could be attributed, but must ensure that they do not thereby make their account vacuous.

In sections 2 and 3 I present two apparent problems for the interpretationist in picking out the suitable situations which he wants to use in his theory. As one would expect, the different kinds of interpretationism will need to respond in different ways, and the task for analytic interpretationism is more difficult.

Next, consider more directly what should count as suitable counterfactual situations. One might hope that an account of these suitable counterfactual situations would involve giving necessary and sufficient conditions for a possible world to provide us with a suitable counterfactual situation. However, a result from the previous chapter suggests a complication for this project.

In chapter 4, section 1, I argued that there were reasons to doubt that all the thoughts of a subject across time, and even within a time, could be interpreted

together. This means that a situation which is suitable for the identification of some thoughts may be unsuitable for the identification of others. Situations will not be suitable or unsuitable simpliciter, and likewise conditions may not be favourable or unfavourable for interpretation simpliciter. Rather, there are conditions or situations that are good for interpreting some thoughts of a subject and bad for interpreting others. To return to the example of dreams, the situation in which you wake the subject and ask one series of questions about their dream reveals different thoughts from the situation in which you wake them and ask a different series of questions.

This means that we cannot give necessary and sufficient conditions for a situation to be suitable for the interpretation of *any* thought that a subject happens to have. We might still be able to give necessary and sufficient conditions for a situation to be in the group of suitable situations which will include at least one situation where each thought of the subject can be interpreted along with many others, but not all the other thoughts of the subject. However, we should note that the viability of interpretationism only requires that for any thought, there be *some* possible world where it can be identified, and that this world can be referred to by the interpretationist as part of his theory. Therefore, although we shouldn't rule out the possibility of a precise theory of the characteristics of suitable situations, we need not rely upon it. Instead of pursuing necessary and sufficient conditions which may not exist, I will primarily consider apparent problem cases, and show how the interpretationist might address them (see sections 3 and 4 below).

Even if we cannot give complete necessary and sufficient conditions on suitable counterfactual situations, however, we can place two quite general conditions on them from the start. This will help us to characterize them, but will also present us with a challenge.

The two quite obvious conditions are that suitable situations must not involve unacceptable changes to the interpreter or their subject, and they must not take us too far from what we intuitively mean by 'interpretable' and 'interpretability'.

To see the importance of these conditions, consider the follow cases. First, suppose that at time t x lies asleep dreaming, and that interpreter y sits at her bedside with no idea what thoughts her dreams involve. Suppose also that there is a possible world, w, in which at time t+1 y suddenly acquires the powers of Lewis's fictitious mighty knower, and is then able to call on a huge amount more evidence and attribute all of x's dream thoughts (and indeed, all her other thoughts, past present and future) to her. This possible world doesn't provide us with a suitable counterfactual situation. The reason I have argued that the interpretationist should use the notion of PB-interpretability, rather than some of the other notions canvassed, is that he needs to avoid distancing the perspective from which thoughts count as interpretable too far from our own perspective. If we allow reference to worlds where interpreters gain the abilities of Lewis's mighty knower, we undermine this aim entirely.

Consider another example: suppose again that at time t x lies asleep, dreaming. Suppose also that there is a possible world, w, in which at time t+1 x suddenly changes so that she both can and does remember all her own thoughts in perfect detail. Three weeks later, x in w is asked what she was thinking at 3:27am three weeks previously, and as a result tells interpreter y what she thought as part of the very same dream that the x of our world had. Even if we accept that this is a genuinely possible scenario, we won't want to say that this shows that x's dream thoughts count as interpretable in this world. The problem here is that the changes suggested in x are too big. Showing that if you improved our memory so drastically you would make it possible to interpret our dreams does not show that we ordinary humans are interpretable.

The challenge that arises from these conditions is then to make it plausible that there will always be a possible world in which thoughts that are hidden in this world can be interpreted, without such unreasonable changes to the situation. This is addressed primarily in section 4, where I suggest that the important issue is to consider *how* a subject is changed, rather than using some undifferentiated notion of

size of change. I also argue that the interpretationist's notion of interpretability is perhaps not exactly what many people would imagine on first hearing the word, but that it resembles an intuitive notion of interpretability in the respects that are important.

In the next section, I turn to a general worry[51] that someone might have about the interpretationist's ability to refer to suitable counterfactual situations. As we will see, it is only a significant problem for the analytic interpretationist, but even he may be able to overcome it.

2. A general problem

Suppose that Rob thinks 'I'd really like a Mars bar' during a logic seminar, but doesn't express this thought to anyone (it being an unsuitable contribution to a philosophical discussion). Suppose he then forgets this thought, as we often do forget such passing thoughts. In the actual world, there isn't enough evidence for anyone to identify Rob's thought. However, the interpretationist may say that it counts as interpretable because there is another possible world in which Tim writes a note to Rob saying 'What are you thinking right now?' and Rob writes back 'that I want a Mars bar.'

But, the opponent of interpretationism may say, there are lots of possible worlds in which Rob and Tim write notes to each other in this logic seminar. In one of them, Rob will think 'I'd really like a Mars bar' and write a note about this to Tim. However, in another he will think 'I'd really like some coke' and so write this instead, and in another he will think 'I'd really like pizza' and write this. How is the interpretationist to select the first possible world as the one which provides the suitable counterfactual situation which reveals Rob's thought in the actual world?

[51] I am indebted to Adrian Boutel for suggesting this issue to me.

Obviously, the suitable counterfactual situation is the one where Rob reveals the *same* thought that he has in the actual world. However, the opponent will say, this is not something that the interpretationist can stipulate, without referring to the thought he is supposed to be giving an account of, and so introducing circularity into the account.

The opponent of interpretationism may then suggest that from within the 'Yes' camp it is much easier to pinpoint which possible worlds give a suitable counterfactual situation where a thought from this world is interpreted. A Yes-camper can say, for example, that a counterfactual situation may show a thought which is hidden in this world to be interpretable because the same internal state which constitutes the thought in this world also exists in the counterfactual situation, in addition to evidence which allows an interpreter in that world to identify the presence of the internal state under the appropriate mental description. The interpretationist, on the other hand, cannot say this, because it would stop him from being able to say that the Availability Claim reveals something philosophically significant about thought. If we adopt the Yes-camper's solution to the problem, that work will be going on elsewhere – in the account of what internal states etc. are needed for particular thoughts and to be a thinker generally.

I think that there are three replies the interpretationist might give to this objection. First, the interpretationist might say that the problem in fact only applies to the analytic interpretationist. The other kinds of interpretationist should indeed avoid talking about the particular internal state which constitutes a thought when picking out suitable counterfactual situations. However, they can simply stipulate that suitable counterfactual situations for showing the interpretability of a thought in this world involve the same thought occurring in the counterfactual situation. As long as this is not *all* they say, they will still be able to tell us what suitable situations will be like, what interpretation depends on, and how the 'interpretability' they speak of relates to what we might intuitively understand by the word. In other words, as

long as specifying that the same thought must occur in the actual and suitable counterfactual situation does not exhaust their account, there seems no reason to suppose that the account will be unilluminating, even if it is not suitable as an analysis. This, then, tells us something about how derivative, dependence and cartographic interpretationists might pick out the possible worlds they are interested in, and something about how those worlds will be related to the actual world. However, if we want to be analytic interpretationists we must seek another solution.

A second option would be to say that in accordance with a standard account of possible worlds, we should evaluate claims about what thoughts a subject is interpretable as having by looking at the closest possible worlds where the antecedent of the relevant counterfactual statement holds. So, for example, take the statement, 'If Tim were to write a note asking Rob what he was thinking, then Rob would write back that he was thinking that he wanted a Mars bar.' This will count as true because in the nearest possible world where Tim writes the note to Rob, Rob reveals his desire for a Mars bar. The counterfactual statements according to which Rob reveals a desire for something else will count as false, because they do not tell us what happens at the *nearest* worlds where their antecedent is true. Thus, Rob's actual thought that he would like a Mars bar will count as interpretable, and he will not count as having and being interpretable as having the many other thoughts that he might have had, but did not have, at the relevant time.

This proposed solution does not make such explicit reference to the thought that is being said to be interpretable. Nevertheless, there are two concerns that we might have about it. First, it relies on the idea that the nearest possible world where the antecedent of the statement above is true will be a world where Rob has the same thought. We may worry that we could develop the case so that this would not obviously be true. The resulting discussion would surely encounter difficulty with the haziness of the notion of closeness of possible worlds, and this problem would make the second concern more pressing. This second worry is that the analytic interpretationist is not entitled to the idea that the world where Rob has the same

thought and reveals it is closer than the world where he has a different thought and reveals it.

If one finds these concerns pressing, one might move to the third option, which is to specify the way in which the suitable possible worlds must be similar to the actual world, without mentioning thought. According to this option, the suitable worlds are ones in which Rob has just the same categorical properties (e.g. of his brain states) up to the point where the interpreter's probe elicits a response that reveals his mental state.

There is a problem that we might raise for this suggestion: since we are calling on the inner, hidden properties of Rob to determine which the suitable situations for interpretation are, we may worry that we are abandoning the spirit of interpretationism. There are two worries buried within this problem: first, the worry that in calling on such states of Rob, we must thereby cause problems for our account of our knowledge of minds; and second, that in calling on such properties we must be admitting that it is really internal states etc. that are important to the nature of thoughts and the mind, rather than actual and potential behaviour and its interpretation.

I think that both of these worries are unjustified. First, consider the knowledge issue. An ordinary interpreter cannot know the categorical properties of their subject's brain, but even if we adopt this account they never need to. It is a tautology that this condition is one that obtains in the actual world (of course the subject has the same categorical properties in the actual world as they have in the actual world). With respect to counterfactual worlds, the interpretationist uses the condition to *stipulate* which counterfactuals are important. No real interpreter is ever in the position of needing to know if the condition obtains and failing to do so. The issue about knowledge just doesn't arise.

Second, consider the worry that calling on internal states and properties of a subject prevents us from insisting on proper interpretationist criteria for the existence of thoughts. This is unfounded, because we do not need to give, nor need

there be, any reduction of the mental to any physical, internal states of subjects, in order for the solution to work. All we need to say is that the categorical properties of the subject *as a whole* should not change, before the occurrence of the thought(s) in question. The interpretationist has already allowed that there may be physical grounds for the dispositions which provide the criteria for whether a person has a thought or not. They may even have said that this is part of our concept of thought (perhaps as part of their account of mental causation). To call on these grounds in such a general way to establish which counterfactual situations are suitable, while continuing to insist that it is the interpretation of the behaviour in these suitable situations which is important with respect to the existence and content of the thoughts the subject has, does not move us into the 'Yes' camp.

We have then, a suggestion for how even the analytic interpretationist may pick out the possible worlds he is interested in, and another suggestion about the nature of counterfactual situations themselves, in the form of a suggestion about their relationship to our world. The 'general problem' seems not to be a problem for anyone. However, this cannot be all there is to say about suitable situations and the ways they can be picked out. To find out more, I turn to another potential counterexample for interpretationism: deception.

3. Deception

Cases of deception are a major challenge for behaviourism: deception is used in the perfect actor counterexamples, which are employed to challenge both the necessity and the sufficiency of the behaviourist account of thought. Whereas other problems faced by behaviourism seemed less applicable to interpretationism, the problem of deception may seem to carry over. The majority of discussions of deception and

behaviourism consider rather outlandish examples.[52] I will begin with an everyday sort of case, but will also need to consider the case of infinite preferences for deception towards the end of this section.

Imagine a person, Alex, saying 'I bought the last round. It's your turn.' She represents herself as believing this, and makes a convincing show of it. However, suppose also that she does not, in fact, believe it. She actually believes that you bought the last round, but she wants you to buy this one too.

Suppose that you try to identify Alex's thought. To start with, you look at Alex's present situation and behaviour. You know that you actually bought the last round, but Alex *appears* perfectly sincere. In fact, she seems so convinced that she bought the last round that you start to doubt your own memory.

Since they endorse at least a moderate form of type B holism, no interpretationist would claim that such a small time slice ought to provide sufficient evidence for the attribution of such a thought. You should also consider Alex's past behaviour. If she is often dishonest about when it's her turn to buy a drink, this may give you reason to doubt her sincerity in the present case. However, suppose that Alex has not been dishonest about such things before, and it is well-established that she has a poor memory, particularly after a couple of drinks.

Still, the interpretationist allows you more evidence: you need not finalise your attribution before the evidence of Alex's future behaviour is in. If Alex confesses that she lied after you have bought the drinks, or she writes about the incident in her diary, or she boasts to someone else about her ability to deceive you, then there will be evidence that a real interpreter can use to attribute her actual thought. Yet, suppose that Alex does none of these things – suppose that she does her best to maintain that she didn't lie, due to concern about what others might think of her, she doesn't keep a diary, and she quickly forgets about the incident. Suppose

[52] See, for example, Putnam (1965), Bennett (1976) and Ellis (2011).

that any interpreter of Alex in the actual world would attribute to her the thought that she bought the last round, but that this is not the right thought.

My interpretationist has not said that an interpreter is always able to identify a subject's thoughts given information about their actual behaviour, only that the subject must be PB-interpretable. And there are plenty of possible situations where Alex's behaviour would have given the thought in question away: even if she did not, she *might* have confessed, she *might* have started to keep a diary that evening, and she *might* have met a friend in the pub's toilet and told them how she pulled the wool over your eyes. However, what makes this case different from, and perhaps more difficult than, the case of dreams is that we have a potential and apparently convincing interpretation of Alex in the actual world, and we need to say why it doesn't tell us what the thought *is* in the actual world, even if it has been formed using what looks like the correct method for attributing thoughts. Why do the possible worlds where Alex acts otherwise show what her thought really is, rather than the actual world? What interpretationist-friendly reason is there for insisting that, in the actual world, Alex doesn't believe that she bought the last round?

I think that there are two options for the interpretationist here. The first, suitable for all stripes of interpretationist, is to say that in the actual world our methods of interpretation have not been fully applied to Alex, and this is why they do not count as revealing her real thought. She may have been tested to some degree. However, we may insist that there must be *some* conditions under which we could definitely elicit a confession (for example through torture, or some other method of persuading Alex that her health and happiness depend upon complete honesty). If you think that such methods could work, you can say that we could gain a confession as long as Alex doesn't assign a relatively infinite weight to hiding the belief in question. You can then say that the actual situation does not reveal Alex's actual thought because it does not involve the best tests, and you can say that the suitable situations for interpretation are those which provide evidence which leads us to the same

verdict about the thoughts in question as the evidence provided through the most reliable tests would suggest. If the most reliable tests would deliver different answers from those we would give in the actual situation, then the actual situation isn't one in which the thought in question is interpretable using only actual behaviour.

The problem with this reply is the possibility that a person might have a preference with infinite weight for hiding certain beliefs. Provided she keeps this preference, the best tests of her in her current state will surely not reveal her true belief, nor indeed her preference for keeping it hidden.

One possibility here is to question the intelligibility of such infinitely weighted preferences. This could be done on the grounds that they would be too irrational for the subject in question to meet the interpretationist's condition of rationality. In response to such a move, some might object that the call on rationality is inappropriate, since the rationality of all subjects has not been established at this stage. This is the response that Ellis (2011) makes in his discussion of whether Davidson can respond to his case of a person whose overriding goal is to deceive anyone who tries to identify his thoughts. It is true that the move could create a problem for the analytic interpretationist, since they may want to argue for the Rationality Claim using the Availability Claim. There is a possibility of vicious circularity if the Rationality Claim also has to be used to defend the truth of the Availability Claim. However, there are other forms of interpretationism and other arguments for the Rationality Claim (see chapter 7).

The success of an appeal to the Rationality Claim also depends, however, on the content of this claim, and on whether infinitely weighted preferences really would entail that a creature wouldn't be rational in the required sense. This seems very much open to question. The idea that suitable situations may at least sometimes be those in which the best tests are performed seems to be a good one, but if the Rationality Claim cannot allow us to rule out infinitely weighted

preferences for deception, it seems that it cannot completely solve the issue with deception.

The second way to respond to our issue with deception is to say that the problem with the actual situation, and the reason why even this good application of our method of interpreting does not reveal the real thought, is that Alex is not being sincere. Nevertheless, her thought remains interpretable because there are possible worlds in which Alex is, at least at some point, sincere (for example with you after you buy the round, or with her friend in the toilet, or with her newly started diary).

This explanation of the suitable situations for interpreting Alex refers to at least one mental state (the state of sincerity), and it may refer to more, even to the thought being said to be interpretable. It is not, therefore, an acceptable suggestion for the analytic interpretationist. Suitable situations are mentioned in our account of interpretability, and since the analytic interpretationist wants to analyse what it is to have particular thoughts and thought in general in terms of this, they cannot refer to mental states here without circularity.

The other three kinds of interpretationism, with their ability to accept reference to mental states in describing the suitable conditions for interpretation, can call upon this condition in their description of suitable situations. They may then be able to accept the possibility of infinitely weighted preferences for concealment: they can say that when someone has a preference with infinite weight for concealing their thoughts, the suitable conditions for interpreting them are ones in which they later change their mind about what is important to them, and decide that it is actually in their best interests to tell or show an interpreter what they were thinking.

In the case of the cartographic interpretationist, I think that this provides a reasonable answer to the issue with deception. However, it should be noted that the case is not so clear with the derivative and dependence interpretationists. They may be able to refer to sincerity in explaining what we may count as the suitable situations for interpretation. However, they also say that there is an analysis, which does not refer to thought, of what having particular thoughts amounts to. They are

supposed to be able to say, for any thought of a being, what makes it the case that the being in question has that thought, using their preferred analysis. The question remains, will they be able to give such an account for deceptive thinkers? It is not possible to answer this question without also considering what the analysis will be. However, we may note that these kinds of interpretationist will face a problem that is quite similar to the traditional problem that behaviourism faces with deception, and the available strategies for saving the theory will be similar. For some suggestions, see Block (1981) and Bennett (1976).

Three kinds of interpretationism, then, could call on sincerity as part of their account of suitable situations. I will now consider one potential problem for, and one limitation of this suggestion. The problem is that this suggestion appears to threaten our ability to give a good account of our knowledge of the minds of others. After all, an interpreter cannot know how their subject would behave if they were being sincere, without knowing whether they are being sincere in the current instance. In response to this, I think that an interpretationist ought to admit to some limitation on our knowledge of other minds. Still, they should say that even if what determines the success of interpretation in the actual situation is inaccessible to the ordinary interpreter, there are situations in which the subject is interpretable by such interpreters, so the Availability Claim remains true. The possibility of successful interpretation remains. Moreover, the claim that we cannot know with certainty what people are thinking (at least sometimes) because they might be lying is a very commonsense idea.[53]

The limitation of this suggestion is that it certainly cannot provide a complete account of suitable situations. First, it is not clear that it can help us to give an account of the important conditions for interpreting someone who is *self*-deceived. It seems that in such a case, a person might sincerely tell an interpreter what they think they are thinking, and yet fail to reveal all the thoughts involved in their self-

[53] See chapter 12, section 2 for some further discussion.

deception. It also seems possible that the rest of their actual behaviour might fail to establish the self-deception and associated thoughts. The same may be true of certain counterfactual situations involving the self-deceived subject. There are ways that we could try to dispel this challenge. We could argue that the self-deceived person *cannot* be completely sincere with an interpreter while their self-deception lasts, or we could add in the condition that suitable situations should involve the subject also being sincere with his or herself. But we could also respond by admitting that situations involving sincerity aren't always suitable situations (indeed, we'll see that this is true in any case in section 4). A second issue is that sincerity doesn't seem to be a necessary condition on a situation being an excellent one for interpretation. We sometimes seem able to firmly establish a person's mental state using their behaviour even when they are being deceitful towards themselves or others. It seems, then, that although sincerity could be an important feature of a suitable situation for interpretation in many instances of interpreting thoughts, it is unlikely to be either a necessary or a sufficient condition on suitable situations.

In conclusion of this section, the issue of deception appears to tell most strongly against analytic interpretationism. The other kinds of interpretationism can all answer the problem to some extent using the suggestions of this section. However, only cartographic interpretationism can rest content with these suggestions, without fearing that deception may cause problems in another part of his theory. We have also obtained two more suggestions about what interpretation depends upon, and what suitable situations will involve: good tests and sincerity.

We now move on to the next potential counter-example: the person with locked-in syndrome. As I said above, this may appear to present us with a case where there are no genuinely suitable counterfactual situations. In addressing it, we will learn more about what the interpretationist's suitable counterfactual situations must be like, and how his notion of 'interpretability' relates to what we might intuitively understand by the word.

4. The locked-in cosmologist

The locked-in cosmologist suffers from total locked-in syndrome: every voluntary muscle in her body is paralysed. However, she has a rich mental life. In particular, she has a strong interest in cosmology and spends a lot of time trying to work out if the universe will continue to expand, or will end in a big crunch. In the actual world, however, nobody realises that the locked-in cosmologist is having thoughts at all, because she has been mistakenly diagnosed as being in a coma. Thus, in the actual world, the ordinary interpreter cannot work out what, or even *that*, the cosmologist is thinking by employing our usual methods.

Cases like this have been used as counterexamples to positions in the 'No' camp, and they initially appear to provide very strong support for the 'Yes' camp. After all, the reason we do not want to deny that the locked-in cosmologist is possible, the reason we in fact want to say that there really are people with this condition, is that when we perform certain sorts of tests we find that their brains are active in just the ways that our brains are active. We may also understand what prevents them from being able to move, and feel certain that this small difference between them and a healthy person cannot prevent them from thinking when there are so many important similarities. We might initially think that this proves that it is the inner stuff, which we can sometimes access using advanced technologies, but which we certainly cannot access using our everyday methods of interpretation, which is really important in determining whether a creature has a mind and/or particular thoughts, regardless of the potential for interpretation.

The interpretationist does not have to accept this account of the significance of locked-in syndrome. They can instead say that brain activity shows that we are probably dealing with a thinker because it shows that there is in fact the potential for the right sort of behaviour, and therefore behaviour and interpretation in suitable

counterfactual situations. The interpretationist should then go on to say something about what these counterfactual situations are like.

One situation in which an interpreter could discover what the locked-in cosmologist is thinking is the situation where the cosmologist is given the appropriate brain scans, where this leads to the right diagnosis, and where the cosmologist is then chosen as a subject for one of the current programs developing neural-interface technology. This technology involves, for example, implanting a device into a subject's brain to pick up certain electrical impulses so that these impulses can be used to control a cursor on a computer screen.[54] Results from such devices provide incontrovertible evidence that there are people with total-locked-in syndrome.

There are potential issues with calling on such cases as indicating suitable counterfactual situations for interpretation. First of all, the situation is brought about by people with knowledge of the brain which far outstrips that of the ordinary interpreter. We might therefore say that there is a very good sense in which our ordinary methods of interpretation are not up to the task of interpreting the locked-in cosmologist, because we lack the knowledge and skills required to set up this interpretational interchange.

However, this first problem is not serious. Once the conditions are in place, an ordinary interpreter *is* capable of interacting with the subject without any special knowledge. The conversation may be a little different from an ordinary conversation: it may be slower, and limited on the cosmologist's side to yes and no answers or to text. However, these are not significant differences, and the average person would be able to adapt to them. It is consistent with interpretationist principles that it should not matter whether ordinary interpreters can understand how to achieve the situations where communication is possible. After all, they do not know all about the many processes involved in allowing ordinary humans to

[54] See for example Orenstein (2011), which reports a case where a woman moves a cursor by 'intending' to move her paralysed hand.

communicate, like the ways the vocal chords or nervous system work. Part of the point of interpretationism is that it allows us to know about someone's mind without knowing these sorts of things.

There are, however, more serious problems for the proposal. A second problem is that the situation in which we give the cosmologist a neural-interface involves a substantial change to her. Indeed, it is a truly life-changing change, from her point of view. The same is true of another apparently suitable situation for interpreting someone with locked-in syndrome: the case where the subject recovers the use of her muscles, and then tells us what she was and is thinking in the ordinary way. Both cases involve a change in the cosmologist's capacities. We may therefore worry that they change her too much, given the considerations in section 1 above. In other words, we may worry that the cases of neural-interface or recovery are not relevantly different from the case of augmented memory, and so that they should be ruled out on the same grounds.

There are differences we could point out between memory augmentation and having a neural-interface. For example, the latter involves restoring an ability to the subject, whereas the former involves augmenting an ability. However, if we depend upon this difference, it may seem that we cannot allow a case where a subject has always been both aware and paralysed.[55] Perhaps we could still persist in this line of response, however, by noting that in the second case we give an ability which is normal for the kind to which the cosmologist belongs, which constitutes a move in the direction of *proper functioning*, whereas in the memory augmentation case the subject gains superhuman abilities. We might then question whether we can make anything of the notions of normal abilities, proper functioning and the idea of a subject belonging to one particular kind, so there is further work to do in order to pursue this response. Nevertheless, it seems to have some potential.

[55] Such cases are mentioned in Child (1994: 29).

Another option would be to argue that, unlike the memory case, using a neural interface or recovering from paralysis are not changes to the *mind* of a subject (though they may be changes to her brain). Although they change her abilities to express her thoughts and produce behaviour, and may thereby change what she goes on to think, they do not change her capacity to have particular thoughts. Memory augmentation, on the other hand, changes a distinctively mental capacity of the subject.

This suggestion about suitable situations raises two issues. First, if we specify that suitable counterfactual situations should not involve a change to the *mental* abilities of the subject, then we once again risk circularity in an analysis of thought. As with sincerity, this is a problem primarily for the analytic interpretationist, and it may simply show that the analytic interpretationist should not take this route. It also raises an issue for the derivative and dependence interpretationists: although they can use this sort of condition to describe suitable situations, they cannot use it in their analysis of why she counts as having the thoughts she does.

The second issue is more significant: it is that in admitting that the change of situation changes the subject's ability to express herself (as we surely must, whatever our account of suitable situations), we seem to be in danger of undermining the idea that she is, in the actual situation, really interpretable. Indeed, this response serves to highlight the fact that there is a very straightforward sense in which the locked-in cosmologist, as she was initially described, is not interpretable using our ordinary methods. Whether or not the interpretationist wants to make use of a distinction between mental and other abilities, they need to say something about this.

This brings us to the difference between two terms that I have used: *potential* and *ability.* I have said that when we discover brain activity, we realise that there is the *potential* for behaviour and therefore for interpretation, and I've admitted that the locked-in cosmologist doesn't have the *ability* to produce behaviour.[56] As I use the

[56] The distinction I draw here is related to Block's distinction between dispositions and capacities in his (1981), and will serve one of the same purposes (namely showing why cases involving paralysis

terms, a person has an ability to do something if either she does it, or she could do it given only the addition of the starting condition that she decides to. A person has the potential to display behaviour relating to thought x, and therefore to be interpreted as thinking x, if they have the ability to display such behaviour, or they would have the ability to display such behaviour if they somehow gained a working output mechanism. It should be noted that calling on the idea of a working output mechanism does not move us into the 'Yes' camp. In the first place, such a mechanism is not required for thought, only used to specify those conditions in which thoughts can actually be interpreted. Second, the notion can be employed by the ordinary person or interpreter without specifying how such a mechanism might work or what physical states are involved.

Since the notion of interpretability that is important for the interpretationist must be linked to the notion of potential, we must then admit that when the interpretationist says that an interpreter could interpret any genuine thought, they are using a specific and somewhat attenuated sense of 'could', requiring less of a subject than one might naturally suppose. Nevertheless, this notion of interpretability seems coherent, and is clearly *related* to what we might intuitively understand by 'interpretability'. Moreover, the sense in which someone like the locked-in cosmologist counts as interpretable does not remove us further from our ordinary practices of interpretation: in the suitable situations, interpretation can proceed much as usual.

This account once again seems unavailable to the analytic interpretationist, since the definitions of ability and potential I have offered refer to mental states. The analytic interpretationist must come up with a different way to explain the sense in which we can still count the locked-in cosmologist as interpretable, perhaps simply by saying more about what counts as a suitable situation, and showing that the cosmologist is interpretable in suitable counterfactual situations. The dependence

---

need not be counterexamples to theories within the 'No' camp). The distinctions do not coincide exactly, however.

and derivative interpretationists can use the distinction in their account of suitable situations and interpretability, but cannot use it in their analysis of thought.

Like the previous section, this section has explored ways that the interpretationist might respond to a potential counterexample, and I have argued that the interpretationist can indeed respond to the case of the locked-in cosmologist. I have then used this discussion to suggest some more potential conditions on suitable situations for interpretation. This time, I have suggested that it must be certain kinds of changes to a subject which are disallowed in suitable situations, rather than just 'big' changes. For example, we might disallow situations where the subject gains superhuman capacities, but not ones where she has ordinary human capacities (such as for communication) restored to her; or we might allow changes to the physical capacities of a subject, but not ones to their mental capacities.

The suggestions for conditions on suitable situations require further elaboration, and it is once again noticeable that fewer options are open to the analytic interpretationist. Nevertheless, possibilities for developing an account of suitable situations are available, and we can argue that another apparent counterexample to interpretationism no longer looks so worrying once we adopt the right account of interpretability.

5. Conclusion

Given the understanding of interpretability argued for in chapter 3, there are ways for the interpretationist to respond to apparent counterexamples involving deception and paralysis. Looking at such cases also helps us to explain what we might mean by saying that a person would display appropriate behaviour *in one suitable counterfactual situation*, and to explain why the actual situation doesn't always allow us to interpret a subject. This in turn helps us to understand what

interpretation depends on and how the interpretationist's notion of interpretability relates to what we might intuitively understand by the word.

I have suggested several principles which may help us to pick out suitable counterfactual situations, but I have not developed them fully or tried to weight them and combine them into a precise theory of suitable situations for interpretation. This might be an achievable task, but it might rather be the case that we cannot give necessary and sufficient conditions, and that the best account takes 'suitable situation' as a cluster concept.

Another thing which the discussion above suggests is that there are far fewer options available to the analytic interpretationist who is trying to give some sort of account of suitable situations, than to other interpretationists. We have not gone into the issue in enough detail to show that the analytic interpretationist cannot give the account they need. However, the prospects for analytic interpretationism here do not look particularly good. It is perhaps therefore the least promising of the interpretationist projects available.

The dependence and derivative interpretationists, while probably more able to give the account they require of suitable situations, are also seen to have a additional challenge in responding to cases of deception and paralysis. Again, I have not shown that this challenge could not be met. Still, the absence of the need to meet such additional challenges may be a significant advantage of cartographic interpretationism, as long as we can justify the claim that it is still a respectable and interesting philosophical account of the mind.

In the next chapter we turn to two more influential examples which are supposed to show why behavioural dispositions and/or interpretability cannot be sufficient for thought. I will argue that the interpretationist should respond by accepting the possibility of certain strange thinkers.

Chapter 6 – String-Searching Machines and Martian Marionettes

In this chapter we consider whether there could be a thing which was interpretable as having thoughts, but which nevertheless did not count as a thinker. Many philosophers have insisted that there could be such a thing, and have offered this as a decisive refutation of views in the 'No' camp. The most famous examples used here are the string-searching machine from Block (1981) and the Martian marionette of Peacocke (1983). These are introduced in section 1, and form the basis of my discussion.

Child (1994) offers a list of potential replies to such examples. However, some of them involve rejecting the sufficiency of interpretability and adding other criteria to our account of thought. As explained in chapter 2, rejecting sufficiency amounts to rejecting interpretationism as I have characterised it. Therefore, I do not pursue these suggestions, but concentrate on the three replies suggested by Child which are open to my interpretationist. Section 2 considers the potential of rejecting the thought experiments on grounds of physical impossibility; sections 3 and 4 explore denying the status of thinker on interpretationist-friendly grounds, in particular because of the history of the supposed thinkers (a suggestion also developed by Ben-Yami (2005)); and section 5 explores how we might accept the things in question as thinkers, and then try to dissipate or discredit the common intuition that they are not (a strategy also offered in Dennett (2009)). I argue that the third of these responses is the most promising, but that Child's form of the response is far more helpful than Dennett's. I then develop Child's account further.

1. The thought experiments

1.1 The string-searching machine

In his (1981) paper, 'Psychologism and Behaviourism', Ned Block's strategy is to describe a machine which can respond sensibly to inputs, and which is therefore intelligent according to a test like Turing's, but for which 'knowledge of the machine's internal processing shows conclusively that it is totally lacking in intelligence.' (19) 'Intelligence' here means the capacity for thought.

The machine that Block describes has sensible strings of conversational contributions recorded inside it. When it receives a conversational contribution, A, from its interlocutor, it selects all those strings beginning with A. It then picks one of these strings, and delivers the next conversational contribution in that string, B. The interlocutor delivers another contribution, C. The machine picks out all those strings which start with A followed by B followed by C, selects one and then delivers the next conversational contribution, D. And so on. All the questions, answers, and other sorts of contribution are contained within the machine before the conversation begins. The only 'moving part' of the machine is the mechanism which identifies the interlocutor's contribution and maps it to the contribution the machine will give.

According to Block, the set of strings of reasonable conversational contributions is a finite set which could in principle be listed by a large team of clever individuals who have been given a large grant to work on the project, who probably have some sort of mechanical help, and who can exercise their own '*imagination and judgment* about what is to count as a sensible string.' (20) Indeed, presumably we should suppose that the machine was created by just such a team.

After describing this machine, Block states that 'the machine has the intelligence of a toaster. *All the intelligence it exhibits is that of its programmers.*' (21)

The thought experiment is initially set up so that the machine only has to pass the test of having a one hour conversation. However, Block claims that no particular

time restriction is necessary.[57] Nor does he think the experiment is limited to cases where inputs and outputs are typed: we can have the machine linked up to a robot body, which delivers 'sensory' information to the machine as input and performs all sorts of behaviours in accordance with the machine's outputs. In this case, the programmers must program a great deal more information into the machine: they must code in not only every sensible opener to a conversation, and every turn the conversation can then take, but also every piece of sensory information the robot body can receive, and a sensible action or set of actions for the body to take for each possible combination of sensory inputs, given all previous sensory inputs received and actions taken. What counts as a sensible action will of course depend upon the robot body's physical capacities, and we may choose whether or not to imagine the robot to be able to exhibit a large range of our behavioural nuances. Block's idea is that *no matter how closely this being approximates our behaviour, indeed even if it is behaviourally indistinguishable from us, it still cannot count as a thinker because of its style of internal information processing.* This is supposed to rule out the possibility of us admitting that the machine lacks thought, but for a reason amenable to those in the 'No' camp. For the majority of this chapter, I focus on this final case with the conversing robot body. I do this because interpretationism as I have characterised it tells an interpreter to consider the environment and actions of a subject, as well as conversational output.

The paper in which Block presents this string-searching machine also briefly mentions another thought experiment which Block thinks delivers the same problem for a theory in the 'No' camp. This alternative machine controls a robot which simulates a particular person's behaviour by using a description of her physical status, her neurobiology or her psychology to work out how she would behave in the situations in which the robot is placed. This robot will be interpretable as thinking if the person it mirrors is, but Block complains that 'it is hardly obvious that

[57] The programmers will, however, have had to program in strings of *some* finite length.

the robot's process of manipulation of descriptions of [her] cogitation is *itself* cogitation.' (41) The issues this case raises are dealt with by considering the Martian marionette case below.

All of Block's thought experiments are supposed to show that the capacity for thought depends not only on the behaviour that can and does result from it, but also on the style of information processing that produces that behaviour. Since in the cases described we are not supposed to be able to determine the nature of the internal information processing by everyday methods of interpretation, we are supposed to conclude that interpretability cannot be sufficient for thought.

1.2 The Martian marionette

In his (1983) book, *Sense and Content*, Peacocke imagines a human body (called 'The Body') which is just like that of an ordinary human, save that it does not contain a brain. The nerves which would normally connect to a brain are instead linked by radio to a computer on Mars. Highly intelligent Martians have designed this computer, have made sure it has and/or collects information about the history of 'The Body', and have given it conditionals specifying what a typical human would do given any history and current stimulation. The computer can therefore make The Body behave in just the way a typical human would in any situation. Thus, The Body is supposed to be interpretable, yet Peacocke claims 'we have a strong intuition that The Body (or The Body plus computer and radio links) does not have propositional attitudes at all: it is just a Martian marionette.' (205)[58]

Peacocke's case is importantly similar to Block's final case above. Although Peacocke adds the details that the computer is very far away and was built by Martians, he does not think that *these* features are the ones that determine our supposed response to the case. Rather, he identifies the problem as follows: 'There

[58] Throughout this chapter, I will refer to the being Peacocke describes as a marionette, even though I will ultimately be arguing that it is not a mere marionette, but a genuine thinker.

are no states of The Body plus computer related as folk psychology takes belief, experience, memory, intention, and so forth to be related.' (206)[59]

Another difference between Peacocke's Martian marionette and Block's final machine is that Block has his machine simulate a particular human, whereas Peacocke says The Body responds just as 'the typical human' does to its history and situation. This difference is, I think, a definite point in favour of Block's case, since we may wonder whether there is any such thing as a 'typical' human response to many kinds of stimulation. I will therefore amend Peacocke's case by imagining that the Martians develop a particular character with its own personality and proclivities.

We have, then, two famous thought experiments which are supposed to disprove the claim that interpretability is sufficient for thought. I now turn to the first way we might respond to these cases: by complaining that they are physically impossible.

2. Physical impossibility

The charge of physical impossibility is one principally directed against Block's string-searching machines. It is an objection that Block himself anticipates and discusses. As Block develops the challenge, the problem is combinatorial explosion: given the huge number of possible conversations, let alone overall interactions of a body with the world, even the machine which is only supposed to cope with an hour long conversation needs to search through more strings than we currently believe that there are particles in the universe. Therefore, no machine could store all the information that Block says his string-searching machines contain, and no group of humans could possibly have programmed them. The response then goes on to say that, because the case is physically impossible, interpretationism does not need to

[59] In diagnosing the problem with the marionette, Peacocke thus seems to gesture towards the very issue about mental causation that this thesis has set to one side. I will therefore not be able to deal fully with Peacocke's complaint. My focus will be on seeing if Peacocke's intuition about the marionette can be explained in another, interpretationist-friendly way, or else discredited.

consider it, and it cannot be used to refute any view of the mind. Although this argument is directed at the string-searching machines, it seems likely that a similar argument could be developed for the computer which is supposed to contain conditionals telling it how a body with any history would react to any stimulation, including stimulation in the form of conversation with other people.

Block's primary[60] answer to this problem is that, since something like the interpretationist's Availability Claim is supposed to be a conceptual claim, we need only a logically, not a physically possible case in order to refute it. I think that this response chimes with the spirit of interpretationism. After all, if our concept of thought is separate from our ideas about the physical processes etc. that go on inside the thinking creature, as interpretationism suggests, why would it not also be separate from our ideas about what is physically possible? The insulation of our ideas about thought from ideas about how thought might be realised should not be selective without some justification, and none has as yet been offered. I therefore do not think that the interpretationist ought to use this response.

Nevertheless, the point about physical impossibility might be used in a different way: we could say that a strong intuition that there couldn't be such a thinker, because they are physically impossible, somehow interferes with our reaction to the case we are being asked to imagine. To give this response would be to move to the strategy of section 5. Before considering it, I consider two ways of trying to disqualify machine and marionette from being thinkers by saying that they have the wrong sort of history.

3. Inappropriate histories: option 1

Ben-Yami (2005) offers a defence of a view called *criteria* behaviourism. This says that psychological concepts have behavioural criteria. According to Ben-Yami, 'no

[60] Block also questions whether his machine really is a physical impossibility; see (1981: 32).

matter to what extent two persons differ in their internal processes, if these differences would not show in their behaviour, then these persons are psychologically the same.' (180) This is something the interpretationist agrees with, provided that 'difference in behaviour' is understood broadly enough. Although Ben-Yami does not call himself an interpretationist, his view shares the feature which makes the thought experiments in this chapter problematic, and it is therefore worth looking at his response to the problem.

Ben-Yami agrees with Block that the string-searching machine (often called Blockhead) lacks intelligence, but says that this is not due to the nature of its internal information processing, but because it lacks the requisite capacities.

The argument uses a comparison with the case of Cyrano helping Christian to give poetic answers to Roxane in *Cyrano de Bergerac*. In Act III of this play, Cyrano hides beneath Roxane's balcony and whispers to Christian what he should say. Although Christian speaks the poetic answers, and Roxane attributes them to him, the fact that he is only repeating them undermines any attribution of poetic ability to Christian. In just the same way, the fact that all Blockhead does is to reproduce the answers it has been given shows, according to Ben-Yami, that it has no intelligence of its own.

The comparison is supposed to show that, like Christian, Blockhead has limited capacities: it can give intelligent answers 'only in *exceptional circumstances*, when it has been programmed in a very specific way. His machine does not have any independent intellectual capacity, but has to be given the right answers in advance.' (185) The programmers might easily have made mistakes, or decided to enter silly answers, and there would have been nothing Blockhead could have done to compensate.

Ben-Yami then suggests that Block's problem was that he focused on the point of view of an interrogator in a Turing test situation:

> [T]he interrogator cannot, indeed, distinguish between the machine and an intelligent person. But the interrogator does not have all the information required to determine the machine's capacities. To determine that, we need to know how the machine interacts with its environment in other circumstances as well – we need to know what happened to it in the past, and not only what it can do at present. (185)

The last sentence above is particularly important for the interpretationist. If Ben-Yami had supposed that Blockhead's lack of capacities was undetectable by the ordinary interpreter, then his reply would not have been interpretationist-friendly, since the machine would have been interpretable in the sense important to the interpretationist and yet would not count as a thinker. However, Ben-Yami suggests that knowing the history of Blockhead would allow us to establish that it does not have the right capacities, and this is something an ordinary interpreter *could* know.

To evaluate Ben-Yami's response, I will consider the analogy between Christian and Blockhead. I argue that there are several relevant differences.

A first difference, which Ben-Yami recognises, is that Cyrano supplies responses to Christian in 'real time', whereas Blockhead has been programmed in advance with all possible sensible responses. This is a point which Block thinks is important. Block compares his machine with the case of a two-way radio. Such a radio can also emit sensible responses in a conversation. However, Block thinks it lacks a capacity that Blockhead has. He thinks we can see this by noting that causal signals from the conversation partners do not reach the programmers, whereas the person on the end of a two-way radio does hear questions and think up answers. The causal efficacy of Blockhead's programmers is therefore limited in a way that allows Blockhead to have a capacity which the radio lacks.

Christian seems to be rather more like the radio than like Blockhead in this respect, and Ben-Yami should admit that the machine has a capacity which Christian and the radio lack: it can produce responses without them being fed into it as the

conversation progresses. However, Ben-Yami must think that this is not the capacity that matters. He must think that Christian and the machine are still importantly similar because, although the programmers don't stand beside the machine during the conversation, it is still similarly dependent on them to produce any answers that it gives. We should then employ Ben-Yami's point above: just as Christian would have given silly answers if Cyrano did the wrong thing, so the machine would give silly answers if the programmers programmed it differently. Thus, according to Ben-Yami, the machine's capacity is limited to 'exceptional circumstances'; its capacity is not 'general' enough.

The problem here is that, for *any* thinker, if there is something wrong with whatever it is that supports her mental processes, she will experience difficulties in producing sensible responses to her situation. These difficulties will be of varying severity depending on what is wrong: sometimes they may result in strange thoughts and responses, which nevertheless can still be counted as genuine thoughts and actions; sometimes we may end up with involuntary behaviour alongside genuine actions, and sometimes a problem stops a biological creature from thinking at all. The problems likely to be experienced by the string-searching machine may be different from our problems, both in causes and effects. But what reason have we to think that our capacity for thought is any more *general*?

Perhaps here Ben-Yami would want to call on the immense difficulties in creating Block's machine. Indeed, we saw above that Blockhead might be impossible, whereas we know that lots of humans do successfully think and respond to their circumstances. If the impossibility of producing the machine is the issue, then we must refer Ben-Yami to the discussion of section 2. If mere difficulty is being called on, and hence the likelihood of the programmers making mistakes, then Ben-Yami seems to be proposing that we rule out candidate thinkers just because the circumstances which led to their existence were highly unlikely. This is a feature which many things we want to count as thinkers may share, and there seems to be little justification for the interpretationist to call on it.

Once created, as per the hypothesis, the machine *does* have a pretty stable and general capacity to respond to things sensibly and in a way that makes it interpretable. I therefore conclude that there is not enough justification for saying that the machine does not have the required capacities. The difference which Ben-Yami recognises between Christian and Blockhead is a significant one, despite his argument to the contrary.

A second difference between the cases further undermines the analogy that Ben-Yami wants to draw: poetic ability and intelligence (in the sense at issue) are importantly different. Although Christian definitely doesn't have the former, we don't say that he therefore isn't entertaining the poetic propositions that Cyrano gives to him. Yet it is precisely the mere thinking of thoughts that is at issue in the case of Blockhead. We should, and can happily, deny that Blockhead has certain abilities that we may associate with our everyday notion of intelligence. In particular, it lacks a certain strong sort of creativity[61]. However, in denying it some of the abilities that we suppose humans to have, we need not deny that it thinks. The sense in which Christian is shown not to have a capacity is therefore quite different from the capacity that Ben-Yami wants to deny in Blockhead, and the analogy may therefore fail to support his case.

Suppose we accept the above criticisms of Ben-Yami's argument. We might still feel that Ben-Yami's central concern is important. As well as presenting an argument that Blockhead lacks a capacity, he also points out that whatever responses Blockhead makes, and whatever capacities he has, these depend upon the intelligence of others. The machine's responses are derivative: parasitic upon those

[61] Blockhead appears to lack what Boden (1994) calls historic creativity, which is the ability to have ideas that no one has ever had before. It need not lack what Boden calls psychological creativity, which involves producing ideas which the subject could not have had before, nor need it lack the ability to have ideas which it merely happens never to have had before. That Blockhead can possess these latter kinds of creativity will be a consequence of the arguments of section 5. We may think that humans are sometimes creative in the first sense, and perhaps this is important in an account of the value of artworks. However, it is surely the latter two kinds of creativity which should be important in the philosophy of mind.

of its programmers. This is the central problem that Ben-Yami sees with Blockhead. He uses it to attempt an argument that Blockhead lacks the relevant capacities, which I have argued fails. But maybe this dependence could be problematic in another way. To consider the nature of this dependence, and the way in which it should affect our interpretation of Blockhead, I will again compare the case of Blockhead with that of Christian.

In the play, Cyrano *tells* Christian what to say. Christian listens to him and purposefully repeats exactly what he is told. I said above that we need not deny that Christian entertains the propositions suggested to him, but nor need we insist that he does so. We *could* imagine that Christian does not even understand the things that Cyrano tells him to say. In this situation, there is an intentional characterisation that we can give of Christian according to which he definitely doesn't generate the content of his responses himself, in any sense, and which even tells us that he doesn't understand them. This intentional characterisation *of Christian* tells us that his responses are derivative in a sense which matters to poetic ability.

In the thought experiment, however, Blockhead has responses *programmed* into it. It does not listen to its programmers, nor decide to use their responses. It need not trust its programmers, nor have any view on them at all. It need have no awareness of the responses as *given* to it. Unlike in Christian's case, there does not seem to be an intentional characterisation of this very machine which tells us that its responses are problematically derivative. The problem, if there is one, must be in giving an intentional characterisation at all, and so we need a justification for saying that this is impossible or inappropriate.

There remains the fact that exactly the responses that the machine produces were thought of by others beforehand, and that this plays a direct causal role in making the machine respond as it does. But it is unclear why, *on interpretationist grounds*, this should make such a difference. Child (1994) suggests that part of the problem may be that the machine's intelligence is parasitic in the sense that it can only arise if others have been intelligent before: not all thinkers could think in this

way. But it is surely equally mysterious why this should matter. To use this response to the thought experiment, the interpretationist must claim that it is part of our concept of thought that it must not depend in this way upon the thoughts of others. But without a justification for this claim, it is an ad hoc, and therefore unconvincing, addition to the theory. A justification cannot be given using Dennett's conception of the purposes of the intentional stance: Blockhead is just as predictable from the intentional stance as a human. In Part III, I question Dennett's account of the purposes of interpretation. Nevertheless, the alternative I offer gives no more support to the claim that the lack of such dependence is important.

It could also be noted that certain world views make all human thought and action derivative in just the way that Blockhead's responses are derivative. Certain varieties of religion offer an account of God's sovereignty according to which God has predestined every aspect of his creation, including the thoughts and actions of his human subjects. Such a view may have consequences for our views on moral responsibility, but surely it would not imply that none of us are thinkers? Certainly, this is not the consequence that the relevant believers draw from their view, even though it makes their thoughts and actions derivative, and even though it means that not all thinkers can be like them, since a God is needed to create them.

I conclude that the interpretationist would do better not to claim that the fact that Blockhead's responses were programmed into it, and therefore that its responses depend upon those of its programmers, means it cannot be a thinker.

It is unclear whether Ben-Yami would want his response to Blockhead to be carried over to Block's other machine and the Martian marionette. Since the response does not appear to work well in the case it was designed for, there is no reason to try extending it to other cases.

4. Inappropriate histories: option 2

The mere fact that it was programmed does not appear to be a good reason for the interpretationist to deny that Blockhead is intelligent. However, there is another way in which the history of Blockhead might make a difference to whether his current behaviour reflects genuine thoughts. This proposal is briefly sketched in Child (1994), and has a precedent in Davidson's discussion of the swampman.[62]

The suggestion calls on the idea that the causal history of a subject, the environment they have been in and the objects they have interacted with all play an important role in determining the contents of their thoughts. Child applies this to the case of Blockhead as follows: 'since his dispositions do not result from any pattern of causal interactions between him and the things and kinds he seems to be thinking about, he cannot correctly be interpreted as thinking about them.' (1994: 45) This can also be applied to Peacocke's marionette, as I show below. The argument mirrors Davidson's reasons for denying that a perfect replica of himself which behaved exactly like him, formed by chance when lightning strikes a rotten log in a swamp, could have thoughts.

I am going to go through a series of objections to this proposal which I hope will show how difficult it would be to endorse. The final objection then shows that it cannot prevent the interpretationist from needing to accept at least *some* very strange thinkers.

First of all, we might question whether Davidson is right about the swampman. We might argue that, given the best account of interpretationism, the swampman should be credited with thought just as soon as he appears in the swamp and starts acting just like Davidson.[63] If this argument worked, surely it would stop us from saying that Blockhead and the marionette can't be thinkers because they haven't interacted

[62] See the beginning of 'Representation and Interpretation' in his (2004).
[63] My hunch is that this is correct, but I will not argue for this here.

in the appropriate way with the things they talk about. However, for the rest of this section, let us suppose that Davidson is correct that the swampman would not immediately be a thinker.

We should then ask what we ought to say about the time after the swampman has cleaned himself off, walked into town, and lived Davidson's life for a few years. He will then have interacted with things and people in the normal way over a decent length of time. Will this mean he has become a thinker? I'm not sure that Davidson wants to continue to deny that the swampman is a thinker as the swampman grows older, but I have not found an explicit ruling on this issue from Davidson. If the swampman can become a thinker, then maybe Blockhead and the marionette can too. After all, they too will interact with the people and things around them. Nevertheless, let us suppose, for the sake of argument, that the swampman never would become a thinker. I will show that, even then, we will still fail to rule out all Blockheads and marionettes from being thinkers using the argument under consideration.

Before that, we should consider a feature peculiar to Blockhead, which might still prevent us from denying it the status of thinker using the argument under consideration. This is that, in Blockhead's case, there *is* a causal connection between the things talked about and the responses and thoughts of the being under consideration: it comes via the programmers. All those things which, when first switched on, Blockhead is disposed to talk about, the programmers must be appropriately causally connected to. Otherwise, they would not have been able to program in the right responses.[64] The connection between Blockhead's thoughts and the things about which it seems to think is admittedly non-standard. Importantly, Blockhead itself may not have interacted with the things it seems to think and talk

[64] Unless they too are Blockheads, or Martian marionettes, which of course are supposed to be able to do everything (non-intentionally described) that a normal human can do, presumably including programming. But a chain of Blockheads will presumably end in real human programmers eventually. Or if the programmers are Martian marionettes, then that is the case we should be discussing, as below.

about before seeming to think and talk about them. However, if the constraint of causal interaction is to be justified by the *use* of noting and calling on such interactions in the process of interpretation, then the way in which the connection is strange does not seem to mark an important difference. If the interpreter could know about the programming of the Blockhead, and about the interactions of its programmers with their environment, this might well enable him to give an interpretation of Blockhead which is useful in just the ways that interpretations are supposed to be useful. Thus, the constraint of appropriate causal connections does not necessarily help with the standard Blockhead case.

Maybe we would find a way around this obstacle to denying that Blockhead thinks if we looked more closely at the sorts of causal connections and interactions with the world that matter to content. *Perhaps* a full account of these would give us a way to discount the causal connections Blockhead has, via its programmers, to the things it talks about, without ruling out the possibility of beliefs formed through testimony and innate ideas. Let us again suppose that this is the case, for the sake of argument.

We may then also take comfort in the fact that the constraint of appropriate causal connections can be applied much more easily to the Martian marionette. The computer which controls The Body was programmed by Martians, and if they have not visited Earth, they may not have interacted with the things that The Body seems to think and talk about. The Martians' program works using their knowledge of neurophysiology, not knowledge couched in terms of how people react to the external situations which they conceptualise in a particular way. In the standard Martian marionette case, then, the causal connections which are supposed to be important may well be missing.

Nevertheless, we can adjust the thought experiments so that there will still be problems. In the case of the Martian marionette we can adjust the thought experiment so that The Body definitely does interact with its environment in the right way before talking and thinking about the things in that environment. The

Body can start out as an almost-normal body of a human baby. It can then appear to move around in and learn about its environment and thereby learn to talk and think in increasingly complex ways. It can do this while still being controlled by a computer which works just as Peacocke describes.

The same is true with Blockhead. Since in the case we are interested in it is connected to a robot, perhaps its 'body' will not grow. But in other respects it can be programmed to react to the world just like a growing child. All its responses will still be programmed in ahead of time, but it will also be true that the dispositions Blockhead has at any particular time will result from just the right pattern of interactions between it and the world.

I suspect that at least some of the claims accepted for the sake of argument above would turn out, on closer inspection, to be false. The discussion above therefore offers us several ways to challenge the argument against Blockhead and marionette thought under consideration. To endorse this response to the thought experiments at all, a great deal more work would need to be done. However, even if the claims were true, the final objection shows that the response of this section still leaves us accepting the possibility of some rather strange thinkers with very different internal information processing procedures from the ordinary human. I suspect that both Block and Peacocke would claim that they still have a strong intuition that even the Blockhead and marionette which have been babies cannot turn into thinkers. If so, then we are still going to need to use the strategy below, of accepting the existence of strange thinkers and then discrediting the intuitions against them, for at least some cases.

5. Allowing strange thinkers

The final strategy to be considered is that of claiming that the beings in the supposed counterexamples *are* thinkers. They will not then refute the claim that

interpretability is sufficient for thought. The fact that some people find this highly implausible must then be addressed, for example by trying to mitigate the intuition that such beings can't be thinkers, or by casting doubt on the value of such intuitions. I first of all consider and criticise a recent strategy of Dennett's for achieving these things, before endorsing a strategy suggested by Child (1994).

Daniel Dennett thinks that we should say that Blockheads and marionettes can be thinkers. In the case of the marionette, he thinks that when we imagine the various ways the fantasy could be fleshed out, we will realise that there is no counterexample here. In his (2009) he says:

> If the off-stage controller controls this body and no other, then we were certainly *right* to attribute the beliefs and desires to the person whose body we have surgically explored; this person, like Dennett in 'Where am I?,'(Dennett 1978) simply keeps his (silicon) brain in a non-traditional location. If, on the other hand, the Martian program has more than one (pseudo-) agent under control, and is coordinating their activities (and not just providing, in one place, *n* different independent agent-brains), then the Martian program *itself* is the best candidate for being the intentional system whose actions we are predicting and explaining. (The Martian program in this case really is a puppeteer, and we should recast all the *only apparently* independent beliefs and desires of the various agents as in reality the intended manifestations of the master agent[)]. (2009: 346-7)

However, Peacocke explicitly mentions Dennett's story, 'Where am I?'. He says that the brain in Dennett's story, which is connected to a body by radio transmitters, can indeed think. We may presume that he also supposes that the computer copy of the brain from the story genuinely thinks, since it is said to be 'a computer duplicate of [Dennett's] brain, reproducing both the complete information processing structure

and the computational speed of my brain in a giant computer program.' (1983: 319) Peacocke is quite explicit that he thinks the problem with his marionette is that it doesn't have states related in the right way to be mental states, because of the *way* it generates behaviour. This marks a difference between the computers in Peacocke and Dennett's stories, and it is one that Peacocke thinks is significant.

Dennett does not add much detail to Peacocke's story, and ignores the feature of the story that Peacocke considers most important. If Dennett really thinks he has given the story all the fleshing out we need in order to see that it is not a counterexample, I take it that he just doesn't share the intuition that the marionette wouldn't be a thinker. As it happens, I share Dennett's intuitions here, but in the course of this section I try to do more to recognise and account for the fact that others do have the same intuition as Peacocke.

Next, consider Dennett's response to Blockhead. Dennett suggests the following imaginative exercise:

> Suppose we discover that Oscar Wilde… lay awake nights thinking of deft retorts to likely remarks and committing these pairs to memory so that he could deliver them of and when the occasion arose 'without missing a beat.' Would this cast any doubt on our categorization of him as an intelligent thinker? (2009: 348)

He thinks that this case then shows us, by analogy, how we should respond to Blockhead:

> Just as Peacocke's puppet does its thinking in a strange *place*, this one does its thinking at a strange *time*! (2009: 348)

However, this response seems very unsatisfactory. The scenario which is supposed to help us seems quite inappropriate: unlike Oscar Wilde, there is a very good sense in which Blockhead does not generate its responses all by itself. In section 3, I suggested that we should admit that Blockhead lacks a kind of creativity (historical creativity) which we can still attribute to Wilde.

Moreover, it is unclear *when* Dennett is suggesting that the machine did its thinking. One option is that he is saying that the machine itself doesn't think, only its programmers do, and that we are really interacting with the programmers when we interact with the machine. However, as Block points out, none of our inputs ever causally affect the programmers. In addition, we do not want to attribute to the programmers all the thoughts that we would otherwise attribute to Blockhead: they did not think, 'Oh, what a beautiful sunrise,' or 'This referee's a joke,' or 'I can't believe I aced my exams', or whatever. Unlike Blockhead, they were not in the right situations to have such thoughts.

Perhaps Dennett is instead suggesting that the programmers and machine are a composite thinker, in which thinking done earlier results in thoughts at later and appropriate times. Certainly, thinking can work like this. The thinking Oscar Wilde does late at night may produce particular new thoughts involving indexicals during social interactions. However, since the machine has been designed to act just like a normal person, it plus its programmers would make a highly psychologically disjointed thinker. As suggested in section 3, Blockhead doesn't need to have any notion of his origins or how his mind works. Wilde will probably remember that he stayed awake thinking up responses, and be pleased that he put in the time when it allows him to deliver just the right retort. Blockhead, on the other hand, may perfectly sincerely assure us that he has *just* come up with a new idea, or worked out the answer to a problem, or whatever. I think there is no more reason to count Blockhead and certain time slices of its programmers as a composite thinker than to count a time-slice of a hypnotist and the person into whom he implants ideas as a composite thinker.

Through its implausibility, I think that Dennett's response to Blockhead points us towards one of the reasons why people might mistakenly intuit that Blockhead cannot be a thinker. The reason is another of those canvassed in Child (1994).

The suggestion is that the intuition that Blockhead isn't a thinker comes from sleight of hand with the notion of a lookup table. A normal human might use a look-up table to produce behaviour while not having any of the thoughts that that behaviour implies. Perhaps people think of this sort of case, and then retain the idea of consciously consulting a lookup table when imagining a being which produces *all* its behaviour like this. They then carry over the intuition that such behaviour wouldn't be intelligent. However, this gives the wrong picture of what is happening with Blockhead: it doesn't literally consult the table at all. No one does, in fact. As Child says, the operation of Blockhead's look-up table occurs at a sub-personal level. Child then suggests that '[t]he fact that Blockhead's *sub-personal* organisation is structurally isomorphic with a look-up table is irrelevant to the question whether, at the *personal* level, Blockhead has propositional attitudes.' (1994: 44)

We therefore have a potential diagnosis of where the person who says that Blockhead can't be a thinker may be going wrong: their intuition may be arising from thinking of the case in a way which is either incoherent or unjustified. It will be incoherent if they are trying to imagine *both* that Blockhead is aware of an internal process, *and* that it isn't a thinker. More charitably, we might take them to be supposing that *if* Blockhead was a thinker, then he would have to be aware of looking things up on a table, but that he then wouldn't count as a thinker after all. In the second case, the supposition is unjustified because there are many subpersonal processes which go on in our brains which we are not aware of, and it is unclear why Blockhead shouldn't have such processes too.[65]

[65] Perhaps at this point, Block will say that even if we allow that the string-searching machine could simulate some subpersonal processes, the interpretationist also needs to say that it simulates some personal ones. He might then claim that since the string-searching machine *only* searches strings;

I will now offer some support for the idea that both Block and Peacocke are indeed making this mistake. When considering his final machine (which works much as the machine that controls The Body in Peacocke's story), Block offers the following as an 'intuition massage' to encourage us to agree that the machine is not thinking: 'substitute for the description-manipulating computer in [the robot's] head a very small *intelligent person* who speaks only Chinese, and who possesses a manual (in Chinese) describing [the relevant] psychological mechanisms.' (1981: 42) Block points out that the person inside need not know anything about what is happening outside, and concludes 'It seems that the robot simulates… thinking… without itself thinking those thoughts.' (1981: 44)

To this intuition massage, we may reply that just because the robot is thinking, this gives no reason to suppose that a miniature person who was aware of and controlled the manipulations inside the robot should be aware of what is happening outside, nor that it has to think the thoughts we attribute to the robot. Such a little person would not be the one we were interpreting. Rather, he would be like the man in Searle's Chinese room argument, who merely contributed to the mechanics of thinking. [66]

Now, if Block's intuition in the case of his second machine is disrupted by the mistaken assumption that this candidate thinker must be aware of manipulating information in the way a miniature person in his head would be aware of such manipulations, then maybe he is making the similar mistake of supposing that Blockhead must be aware of his own information processing mechanisms. This would then explain Block's intuition, but it would not justify it.

Peacocke likewise may be accused of this mistake in his discussion of the Martian marionette. Peacocke complains that the states of the computer that he has

---

there are no other processes going on inside Blockhead which can be the proper person-level processes that Blockhead is supposedly aware of. My reply to this is given below.

[66] See his (1980). Searle, of course, does not agree that the man in the Chinese room merely contributes to the mechanics of genuine thought, but this well known response to his thought experiment is the one the interpretationist should give.

described as causes of The Body's behaviour aren't related in the right way to be mental states like ours. But if they are sub-personal states which the marionette has no awareness of, it is unclear why this is supposed to matter.

As implied by the analogy with Wilde above, even Dennett appears to be vulnerable to this sort of mistake. In the same paper, he talks about Blockhead having the vast number of potential strings of inputs and outputs stored 'in memory'. However, this is not the way the interpretationist ought to think of the situation. Just because the information stored in the machine has a particular meaning for its programmers, this does not mean that the physical realisation of the information has the same meaning for Blockhead. The information the programmers put in is not written in a 'language of thought.' It need not even be written in a language that Blockhead speaks or understands. Blockhead does *not* already believe all the things the programmers have programmed him to believe in the right situations – he must wait for the right situations to believe them. Nor does he have a fantastic number of conversations stored in his memory – he could not list them to us, even if he wanted to. We should say that from his point of view, the responses he makes and the intentions he forms seem to come to him in much the same way as our intentions and responses arrive, sometimes as a result of explicit trains of thought, and sometimes apparently out of nowhere.

There is one problem with this account as a diagnosis of Block's intuition. I have suggested that Block makes a similar mistake to Searle in his Chinese room argument. Yet, Block himself has argued against Searle,[67] and in his (1981) he explicitly says that he is not making the same argument. Whereas Searle used his argument to claim that no process of symbol manipulation could count as thought, Block says 'what justifies us in regarding some symbol-manipulating homunculus heads… as unintelligent is that the causal relations among their states do not mirror the casual relations among our mental states.' (12) However, then the addition of the

[67] See his (1980a).

miniature person into the machine should be irrelevant. Certainly, the fact that the homunculus isn't thinking the thoughts we attribute to the robot should not have been called on. Even if Block's own intuitions do not result from the mistake suggested, it seems that he must be attempting to manipulate his readers' intuitions in an unjustified way.

The point about the causal relations among mental states takes us, as with Peacocke's ultimate diagnosis of the problem with the Martian marionette, into questions about how exactly interpretationism should deal with the causal efficacy of the mental. As explained in chapter 2, this is a huge topic, and I have decided to put it to one side in order to explore the more neglected questions of what the interpretationist could mean by the two central notions of his theory, interpretability and rationality. The most I can do here is to point out that the packets of information that the machine or homunculus manipulates need not be claimed to be the beliefs and desires of the candidate thinker. As Child says, this may all be sub-personal processing. The question of whether these sub-personal states can form the basis for a higher level consisting of genuine beliefs and desires etc. which are related in the right way, and which form the person-level processes that Blockhead and the marionette appear to be aware of, remains open. But we should note that Peacocke and Block, in the papers containing the thought experiments we have been considering, have not given us reasons to suppose that they cannot.

I think the above provides a plausible diagnosis of many people's intuitions about string-searching machines and Martian marionettes. Perhaps some such people might recognise that they have made the suggested mistake, and might then find their intuitions changed or lessened. However, even if they do not, there are ways for the interpretationist to cast doubt on intuitive responses to the cases in question.

One of these ways has already been mentioned: we may point to the probable physical impossibility of the cases, and suggest that people may be confusing an

intuition that the thinker described just couldn't exist, and the intuition that the thing in question, if *per impossibile* it did exist, wouldn't be a thinker.

Another thing we might say is that we humans have a strong record of producing chauvinist intuitions about mental states, particularly before adequate interactions with a candidate thinker. People have denied the personhood or mental capacities of other races, but their concepts of personhood and thought *should* be extended to cover at least all typical adult humans. Some people also have a very hard time accepting that a lump of matter like the brain could produce genuine thought and may be attracted to dualism as a result. However, most modern philosophers would suggest that there is something amiss with such intuitions.

When presented with a really strange sort of thinker, it is unsurprising that many of us initially say that it cannot really be a thinker. But if we were to interact with such a being over a long period of time and to mutual benefit; if it made us laugh just like our friends did, and seemed to get upset at just the same sorts of problems as us, and presented us with good ideas for joint projects; or if it knew just what to say to make us furious, or acted spitefully in revenge for a disparaging comment we made, etc., then maybe our intuitions would change.

6. Conclusion

In conclusion, I think that the interpretationist should say that string-searching machines and Martian marionettes do not provide counterexamples to the Availability Claim, because they are just examples of rather strange thinkers.

PART III: RATIONALITY

In chapter 2, I gave the beginnings of an explanation for why the interpretationist's Availability and Rationality Claims came together: I said that the Availability Claim posited that the thoughts of a creature could be identified through an attempt to understand and explain the creature and its actions in terms of the reasons of that very creature; this, it was suggested, meant that any thinker would need to be rational to at least some degree, for how else could this interpretation take place? In this part of the thesis, I consider in more depth why we should suppose that thinkers must be rational, and what this rationality could amount to.

In comparison to 'interpretability', the notion of rationality has been considered more obviously interesting and problematic by philosophers. It has therefore been discussed at greater length, and the philosophical literature contains many more disagreements on the subject, over such issues as whether rationality is an achievement or a capacity; whether and to what degree thinking creatures must possess it; whether it prescribes only more or less logical relationships between thoughts, or also mandates certain appropriate feelings and reasonable ends; and whether it should be separated into practical and theoretical subsections. Those who have arrived at conceptions of rationality have then gone on to use these conceptions in quite different ways and for different philosophical purposes. It will not be possible to address all of the questions that people have asked about rationality, nor to give interpretationist answers to all of them. The aim of Part III is rather to give a schematic view of the notion of rationality which the interpretationist should use.

Chapter 7 explains the nature of the task the interpretationist takes up in trying to characterise rationality, and shows how rationality fits in to the interpretationist project in more detail by considering potential arguments for holding a Rationality Claim.

Chapter 8 considers a simple, rule-based view of rationality, and uses it to show why facts about human limitations and our apparent propensity towards reasoning errors have been considered problematic for the Rationality Claim. A major problem with this simple view is that it fails to accommodate the idea that our limitations, resources and interests can determine which rules of reasoning it is sensible for us to follow. In chapter 9, I suggest a consequentialist account of rationality which sanctions the use of quick and dirty reasoning heuristics, but over-generalises this good idea and so removes all constraints on the content of the thoughts interpreted.

Chapter 10 then argues that Cherniak's (1986) notion of minimal rationality combines the promising aspects of both previous theories, and shows that it can be adapted so that it is suitable for interpretationist purposes. Thus, I argue that the interpretationist should say that being rational involves getting at least some things right, and exhibiting a pattern in one's reasoning and acting. I raise the question of whether these conditions are enough to ensure that all who meet them are genuine thinkers, with their own perspective to be understood in terms of their reasons, and suggest that this minimal account of rationality does correspond to a very basic notion of a point of view. Other features may nevertheless be added to our account to develop a more full-blooded conception of a point of view and the ability to respond to reasons.

Finally, in chapter 11, I consider types of thought (dream thoughts and imaginings) and a kind of thinking (associative thinking) which initially appear to have little to do with rationality. I argue that these too can be fitted into the interpretationist scheme, but that they show an important way in which interpretation may be explanatorily incomplete. I then outline what this suggests about the relationship between the science of the mind, the philosophy of mind, and interpersonal interpretation.

Chapter 7 – The Rationality Claim

1. The nature of the task

Despite the many disputes about rationality, there does appear to be something which unifies most of the various accounts of rationality: namely, that rationality has positive normative import.[68] Rationality has to do with what is good in thinking, and perhaps also to some extent what is good in acting. For example, Dennett states 'I want to use "rational" as a general purpose term of cognitive approval' (1987: 97), while Heal claims that 'rational' expresses a value-laden notion, 'bound up with our view about… what we should aspire to.' (2007: 403) It seems to be generally agreed that when a philosopher claims that the thoughts of an agent must be in some sense rational, they are claiming that the agent cannot be *too bad* at thinking.

However, this small puddle of agreement is not enough to show that there are not importantly different, but equally reasonable notions of rationality. For example, the idea that rationality has positive normative import does not settle whether all good features of thinking are aspects of rationality, or only some subset. Heal (2007) suggests that

> The need for some notions in the area of 'reason' and 'rationality' are rooted in our ability to engage in discursive and persuasive linguistic exchanges. But because such exchanges can (as Wittgenstein emphasises) be so various, we should expect the notions to come in many versions, shaped by history and culture. (403)[69]

[68] Although even this is not universal; see for example Kolodny (2005).

[69] We might develop this point further by suggesting that although the notion of rationality may have been developed in situations involving linguistic exchanges, this does not mean that it is now applied exclusively within or to such contexts: there are many other activities to which an ability to think may contribute in various ways, and too great a focus on linguistic activity, even in all its many forms, might produce a somewhat restricted understanding of rationality.

She then further suggests that although we (here referring to people in our culture, perhaps specifically the culture of Anglophone analytic philosophy) recognise that qualities such as imagination, tolerance and balance are important in intellectual life, nevertheless our idea of rationality is, for historical reasons, focused most firmly on those aspects of thinking exemplified in deductive proofs in areas such as geometry and arithmetic.

What the interpretationist is interested in is whether there is any concept, which we might reasonably term 'rationality', which can both do the theoretical work the interpretationist requires, and make the Rationality Claim sound at all plausible. The interpretationist, then, does not need to pursue some 'essence' or 'true nature' of rationality; he can agree with Heal that there are different notions available, associated with different cultures and different purposes. He only has to suppose that *one* concept, which we might reasonably term 'rationality', can be employed in a plausible Rationality Claim, and then to say what *that* concept of rationality involves.

The obvious starting point for finding the right concept of rationality for the interpretationist is then to consider the role that a concept of rationality must play in his theory. Section 2 does this by looking at various arguments for the Rationality Claim. Finally, section 3 considers the different requirements that the different forms of interpretationism place on an account of rationality.

2. The role of rationality

The term 'rationality' is used by the interpretationist to refer to that state of a system which is required for the relevant kind of interpretation to be possible. 'Rational', then, is a placeholder for the feature, or at least one very important feature, of a system that makes interpretation possible, and it is an appropriate term to use just because the sort of interpretation in question involves the attribution of reasons. It also, however, plays a technical role in the interpretationist's theory, and as a result

may well diverge from some of our everyday uses of the word. For example, while we often attribute rationality and irrationality to particular thoughts and actions, the interpretationist's *primary* interest is in a property attributable to a whole mind.

In order to get clearer on exactly what role rationality and the Rationality Claim play in interpretationism, it is instructive to consider some arguments for the claim. As explained previously, arguing directly for interpretationism and its central claims is not one of the aims of this thesis. However, these arguments for the Rationality Claim, including the manner in which one of them fails, illustrate some features of how interpretationists do and should employ the notion of rationality, and therefore provide some desiderata for a more detailed explanation of the content of the Rationality Claim.

2.1 An argument from probability

First, I consider an argument which Stich (1990) extracts from Quine's *Word and Object*. I agree with Stich that this argument fails, and so do not think that it should be used by the interpretationist. Still, it is worth looking at because the way in which it fails reveals that the Rationality Claim cannot be an empirical claim. Since my primary concern is how an interpretationist uses their notion of rationality, I won't consider the best way to interpret Quine, a self-proclaimed behaviourist, and I do not claim that the argument below should be attributed to him.

Quine considers the possibility of pre-logical mentality, and from this discussion extracts a maxim: 'assertions startlingly false on the face of them are likely to turn on hidden differences in language' (1960: 59). He thinks this is supported by the common sense view that 'one interlocutor's silliness, beyond a certain point, is less likely than bad translation – or in the domestic case, linguistic divergence.' (ibid.)

To convert these ideas into an interpretationist argument, we should switch the focus from the translation of utterances as silly, to the attribution of attitudes

which are, in some yet to be determined sense, silly. We get a group of maxims which counsel us against attributing silly attitudes and groups of attitudes, and against interpreting a thinker as silly. Let us suppose that someone offers this as an argument for accepting the Rationality Claim.

If offered this argument, we should ask why silly attitudes, and hence groups of silly attitudes, and hence silly thinkers, are supposed to be unlikely. The probability of such things doesn't seem like the sort of thing of which we have basic, a priori knowledge: if this is to be a good argument, the probabilistic claims require some justification. Stich (1990) suggests that they could be supported by evidence about how people actually reason. However, to gain such evidence we need to understand what people are doing and saying, and Quine's principle was supposed to guide this. Thus, Stich concludes that 'it looks like any inductive attempt to support Quine's precept will beg the question.' (1990: 36)

What this shows is that the Rationality Claim, the claim that thinkers are rational in whatever sense is required for interpretation, cannot be an empirical claim; and nor can the claim that any particular *thinker* is rational. Before we have accepted and started to employ the Rationality Claim, empirical evidence about how people reason, or whether a particular system counts as a person, is not available. Claims that thinkers are rational in some other sense may be established empirically, and claims that a particular system is rational may be established empirically, as part of the process of establishing the contingent fact about whether or not that system is a thinker. However, the interpretationist's Rationality Claim is supposed to be a priori.

These thoughts do not deny that any particular proposed silliness in thinking is unlikely. The argument is also not intended to show that empirical facts can have no relevance to the plausibility of a Rationality Claim: I will argue that they are indeed relevant in chapter 9. Now, however, I turn to an undeniably a priori and direct argument for the conclusion that the Rationality Claim is true.

2.2 An argument from requirements for interpretation

The second argument to be considered proceeds from the Availability Claim to the Rationality Claim via the premise that interpretation is only possible if the Rationality Claim is true. It is an argument for the *truth* of the Rationality Claim only if you accept that the Availability Claim is prior in the argumentative structure of the theory. However, it is also an argument for any interpretationist to accept the Rationality Claim, in so far as they wish to remain an interpretationist. This latter use for the argument makes it the most important of the arguments for the purposes of this thesis.

This argument seems to underpin Davidson's application of the Principle of Charity. Davidson starts from the problem of how to identify the meanings of speakers' utterances. In considering this problem, he identifies a circle of meaning and belief: 'If all we have to go on is the fact of honest utterance, we cannot infer the belief without knowing the meaning, and have no chance of inferring the meaning without the belief.' (1984:142) In order to break this impasse, he claims that we should 'hold belief constant' while solving for meaning.[70] This is to employ the Principle of Charity, and Davidson explicitly says this involves investing a person with basic rationality[71]. If a person's language is to be interpretable, then, the Principle of Charity must apply to them and they must be minimally rational.

Davidson's version of the argument, set in the context of the problem of how we can interpret language, only has the potential to secure the rationality of language-using creatures. However, the interpretationist who disagrees with Davidson on this issue can still use an argument with the same structure. Consider the case of interpreting the thoughts of a non-linguistic creature. If any environmental input could result in any belief, and any set of beliefs and desires

[70] And the same sort of strategy is supposed to apply to other thoughts which might affect linguistic interpretation.

[71] See for example his (2001: 211).

could lead to any action, the interpreter would have no way of determining the thoughts of the creature. In both versions of the argument, and regardless of why we think thoughts must be interpretable, the essential point is that interpretation is only possible if there are connections between what a subject experiences and does and what they think. In calling the term 'rationality' a placeholder, as above, we are saying that rationality is just whatever turns out to provide or constitute the needed connections.

Given this role for rationality, we can say a little more about what the Rationality Claim must attribute to thinkers. Rationality is supposed to have a central role in allowing the interpreter to determine the content of a subject's thoughts. Whatever rationality is, therefore, a thinker must have enough of it to enable interpretation of them. Moreover, there must be some restrictions on how rationality is to be realised throughout a system of thought. Since it is involved in determining the nature and content of every thought, it cannot be that one simply needs a lot of rationality in certain parts of the system, but can then allow complete irrationality or non-rationality elsewhere, with no account of how the different sections are connected. For example, it cannot be the case that a person has a set of true and rational beliefs about the domain of physics, and then also has one lone and completely isolated thought that 'all English people enjoy shouting at the top of their voices about their innermost feelings,' where this thought is based on no evidence, and is not accompanied by any other thoughts about what it is to be English, what a person is, what feelings are, etc.

Endorsing this argument, then, has consequences for answering the question 'How far and in what ways can we fall short of what rationality demands while still counting as sufficiently rational?' This question will have great significance as we meet some of the problem cases which are supposed to throw doubt on the Rationality Claim in subsequent chapters. The details of the answer depend on how exactly rationality determines the nature of each thought. The argument will be

developed further in chapter 9, as we consider whether the interpretationist could make use of a consequentialist conception of rationality.

It should be noted that there may be other considerations to take into account in producing a good interpretation, and in deciding whether an object warrants such interpretation at all. Some of these could be counted as further conditions on rationality (for example, constraints on desires and ends, which Dennett seems to take as an aspect of rationality[72]), while others are better seen as constraints on interpretation (for example some form of demand for simplicity or economy in interpretation). In this thesis, however, I restrict my attention to what the interpretationist should say about the connections between thoughts, environment and action.

It should also be noted that if, after all such constraints have been added, there is still indeterminacy of interpretation, then this must be accepted by the interpretationist as a feature that thoughts can have. Again, I will not discuss this, since it does not fit into my project of addressing apparently obvious counterexamples to interpretation and developing the central notions of interpretability and rationality.

2.3 An argument directly from the nature of thought

Rather than considering how thought must be if it is to be interpretable, we could instead take a more 'objective' stance and consider what is required to make a thought the thought that it is, regardless of whether anyone is or could be interpreting it. Such an argument thus involves what in chapter 4 I called type C holism about the mental: the view that the content and/or attitude of any mental state depends upon the set of which it is a part. If this is taken to be plausible independently of the Availability Claim, we gain a different argument for the

[72] See his (1987: 20).

Rationality Claim. Such an argument, however, has much the same consequences for the nature of the sort of rationality required by the interpretationist as does the one above. It will not, therefore, receive further independent discussion.

3. Varying requirements

Many of the ideas introduced in the previous sections concerning the nature of the Rationality Claim and the concepts it uses should be congenial to all of the varieties of interpretationism introduced in chapter 2, section 4. However, the different varieties do not all have the same requirements for a concept of rationality for use in their theory.

So far in this chapter, I have talked about what rationality might involve by talking about thoughts and how they must be related to each other (and to actions and environment). For example, I began by suggesting that the one point of agreement among philosophers on this topic is that the rational has something to do with what is good *in thinking*. This way of talking will be acceptable to the cartographic interpretationist, who after all is not attempting to give an analysis of thought. To use psychological concepts while saying how thought and rationality relate to each other is a sensible way to proceed as long as neither notion or group of notions is to be given priority or explained wholly in terms of the other. However, for the derivative and dependence interpretationists, who attempt to analyse thought in terms of rationality or rational patterns, there is a problem.

The problem is a pervasive one: theories of rationality are generally presented by talking about 'thought' and how thoughts are related to each other, to actions and to environment. The derivative and dependence interpretationists, then, may find it difficult to engage with the literature on the subject. Throughout Part III, I ignore this problem, and engage with some suggested theories of rationality and discuss their suitability for what we might call the general interpretationist project. I do not rule out the idea that my conclusions could be adopted, with some adjustment, by

derivative and dependence interpretationists. However, the cartographic interpretationist will find what I say easier to apply.

4. Conclusion

The interpretationist is not committed to there being an essence or true nature of rationality. However, he is committed to there being a concept which might reasonably be called rationality and which can do the work his theory requires of it; most notably, constraining connections between environments, thoughts and actions enough to make interpretation possible. The next chapter begins the task of saying what this notion of rationality could be.

Chapter 8 – The Standard Picture

I begin the task of finding the best conception of rationality for the interpretationist by considering a simple, bold, and (for the relevant purposes) implausible view of rationality. I argue that the interpretationist should not use this conception in their Rationality Claim.

I take this route for three reasons. First, the conception of rationality in question is not obviously a straw man: at least in his earlier work, Daniel Dennett makes comments which seem to presuppose something very like it.[73] Second, it is certainly a view of rationality which philosophers have attacked while considering the Rationality Claim: Stein's (1996) attack on the idea that we are rational is an attack on the idea that we are rational in this sense; and Heal (2003, 2007), Stich (1991) and Cherniak (1986) all attribute rationality claims involving this conception of rationality to others, and then attack them. Finally, I think that looking at this simple view of rationality will give us the best grip on why certain phenomena are sometimes thought to be counterexamples to the Rationality Claim, what the terrain of this debate looks like, and what we require in a more sophisticated conception of rationality that will be of use to interpretationism.

In section 1, I introduce this simple picture under the name 'the Standard Picture' (following Stein (1994)'s terminology). Section 2 then shows how an interpretationist might use this picture of rationality. In section 3 I present the supposed counterexamples to the Rationality Claim, and in section 4 I argue that although some of these may be less serious than has been supposed, some of them are decisive against a Rationality Claim which uses the Standard Picture. There are many other ways one might attack the Standard Picture and a Rationality Claim

[73] See, for example, his (1971).

which uses it, and some of these are flagged during the discussion. However, the focus of the chapter is on arguing that we humans cannot be considered rational in the sense at issue, and this will point the way to subsequent suggestions for the developing the Rationality Claim. I argue that aspects of the Standard Picture may be useful to the interpretationist if further developed, but the simple picture of rationality presented here is not the one he ought to adopt.

1. Outlining the Standard Picture

Edward Stein has used the name 'the Standard Picture' to refer to the following view of rationality:

> According to this picture, to be rational is to reason in accordance with principles of reasoning that are based on rules of logic, probability theory, and so forth. If the standard picture of reasoning is right, principles of reasoning that are based on such rules are normative principles of reasoning, namely they are the principles we ought to reason in accordance with. (1996: 4)

This is supposed to be a familiar, intuitively attractive picture of rationality. Stein does not attribute it to any particular philosopher, but thinks that it is what many philosophers have had in mind when talking about rationality. It is supposed to be a picture that those who are acquainted with this literature will recognise.

It is not difficult to see how one might come to adopt such a view of rationality. After endorsing the widely shared intuition that rationality has something to do with what is good in thinking, an obvious next question is to ask what good thinking involves. Our best theories about how we should reason seem a good place to look: they just are our best and most sustained attempts to work out what good thinking involves.

The theories explicitly mentioned by Stein, logic and probability theory, seem to involve attempts to discern rules for theoretical reasoning. However, as we will see in the next section, the interpretationist needs an account of practical reasoning as well. How to characterise these potentially different kinds of reasoning is controversial, but I will work with the following simple understanding: theoretical reasoning concerns questions of what is true, and/or what we should believe, while practical reasoning concerns questions of how we should act. Theories which attempt to systematise the requirements of practical reasoning include rational choice theory and game theory.

Our actual accounts of what good reasoning involves are incomplete, and some obviously important areas display a great lack of consensus – for example, there is very little agreement about what (if any) rules govern rational responses to new information and therefore belief revision. Nevertheless, there is some (more or less) uncontested ground. From logic, we might pick out the requirements of consistency and completeness. Applied to belief, we might take these to demand that the rational person believe no contradictions and believe all the logical consequences of his or her beliefs. From choice theory, we might pick out the rule that one not have inconsistent preferences.

I take it that Stein thinks that being rational must involve reasoning in accordance with the completed/perfected versions of these theories, and that it certainly involves obeying those rules we are already certain about. It is then a very intuitive idea that we should try to be rational, and reason in accordance with rational principles.[74]

However, saying that the rational being must succeed in obeying all the rules all of the time would be hopelessly implausible. We all know that everyone makes mistakes sometimes. To attack a theory which denied this really would be to attack a straw man, and there would be little point in the exercise. A theory that is worth

[74] Why exactly we should obey the rules of rational thinking is a matter of controversy. I will not enter this debate here.

considering must not make rationality an all or nothing matter, so that either you obey all rules always and everywhere, and count as rational, or break any rule anywhere or anywhen and count as non-rational. We should therefore add to the Standard Picture the idea that only perfect or ideal rationality requires always keeping all the rules. Our rationality must instead involve some sort of approximation to this ideal. Being more rational can then involve being a closer approximation, and being less rational can involve falling further short.

The most obvious way to understand being an 'approximation' to the perfectly rational beings is as obeying a high proportion of the rules almost all of the time. This, of course, is a very minimal concession. Stein suggests another option: that one can approximate the perfectly rational being by having the capacity to follow the same rules as he does, even while perhaps quite frequently failing to exercise this capacity.[75] This relates to the debate, mentioned in the introduction to Part III, about whether rationality is better seen as an achievement or as a capacity. I argue that the concept of rationality required by the interpretationist must involve some degree of achievement in chapter 10. However, for the purposes of this chapter we may accept Stein's suggestion as a possible way of understanding 'approximation' to perfect rationality. It involves a greater concession than the first way, but is still a very strong requirement, as we will see. There are other, more concessive ways to interpret 'approximation'. I won't be considering them in this chapter, and will not count a view which employs them as a version of the Standard Picture.

Having thus outlined the Standard Picture, there are various ways that we could attack it. One option would be to attack the idea that the notion of a perfectly rational being makes sense (a strategy pursued in Heal (2007)). Alternatively, we could challenge the idea that good reasoning can be fully codified in rules (see Child

[75] See Stein (1996), especially chapter 2.

(1993)). Let us set these issues to one side for the moment, and concentrate on how an interpretationist might employ this picture of rationality, on the assumption that it is coherent.

2. Using the Standard Picture

Suppose, then, that we interpret the Rationality Claim as follows: thought must be rational in the sense outlined by the Standard Picture of Rationality. Call this the Standard Rationality Claim, or SRC. We must then consider how this claim could be used by the interpretationist.

The more promising arguments for the Rationality Claim in the previous chapter claimed that rational relations between the thoughts of a mind are needed to determine the nature and contents of those thoughts. If we adopt the Standard Picture of Rationality along with these arguments, we may arrive at a particular picture of how thought content is determined. Heal (2007) describes the situation as follows: we take our best formal theories of reasoning, and use them to 'conceive of abstract items (propositions) and see that, considered as axioms and theorems, they exist in vast timeless patterns of truth-value relations.' (408) We then say that reasoning in accordance with these formal theories is necessary for our thought, because it is only this that allows us to realise the timeless patterns, and it is only by realising these patterns that we can be said to be thinking thoughts with the relevant propositional content. On this account, realising these patterns just is what thought involves.

It seems that the timeless patterns of truth-value relations will be described by those theories associated with theoretical reason, and that these patterns will need to obtain among the subject's beliefs. It is less immediately clear where the desires and other thoughts of a creature, or a creature's actions, fit into this picture. This is where constraints on practical reasoning must enter. As above, practical reason is often described as the form of rationality which tells us what we should do, given

the things we believe and desire. Without some rational constraints on action, it would be no guide to what the subject believed, and so an interpreter could not tell what part of the timeless pattern of truth-value relationships a subject was instantiating. Or to make a more metaphysical point, in line with the argument from chapter 7, section 2.3, without the right connections between what is done and what is believed, a subject just wouldn't count as performing the actions or having the beliefs in question. Therefore, given the account of the determination of thought content above, we should take it that practical rationality constraints determine what it looks like for a subject to instantiate a pattern described by theoretical reason. In performing this job, practical rationality constraints also allow us to determine a creature's desires and characterise their actions. Theoretical and practical rationality will then both be important, but will have somewhat different roles.

To make this clearer, let us consider how SCR would be used in interpretation, for example of the Analogoids from chapter 1. For simplicity, let us here suppose that the Analogoids are simpler creatures than was suggested in Part I, and do not have a language. Suppose that the Analogoids are octopus-like creatures living in an ocean. Most of the time, they move slowly through the water and ingest tiny creatures near the ocean bed, or float near the surface of the water in the sunshine. However, this behaviour tends to alter when a predator enters an Analogoid's vicinity. There are three such predators. The first, a giant sea spider, is fast, but can only move along the ocean bed. When it approaches an Analogoid, said Analogoid usually moves upwards. The second, a sluggish sort of eel, can go anywhere but moves more slowly than the Analogoids. When it approaches an Analogoid, the Analogoid tends to move quickly in the opposite direction. The third predator, which resembles a shark, is fast and able to swim in any direction, but too large to get in amongst the branches of a certain plant that grows on the ocean floor. When it approaches an Analogoid, the Analogoid tends to move into the shelter of such plants.

A human trying to interpret an Analogoid starts with the ability to see how that Analogoid moves. He may then use the assumption that his subject is obeying the rules of practical reasoning to determine some candidate belief and desire states for the creature. When the Analogoid moves upwards while being approached by a giant spider, for example, the interpreter assumes that this action was the best one to perform given the creature's beliefs and desires, and postulates some beliefs and desires which fit the bill: for example, the belief that there is a predator coming towards it, the belief that it can escape this predator by swimming upwards, and the desire to remain uneaten. The interpreter does likewise with various other actions, and the need to refrain from postulating contradictory beliefs provides some initial constraints on candidate beliefs. Given a set of candidate beliefs, the interpreter then assumes that these form part of a pattern of beliefs which instantiate the patterns of truth-value relations mentioned above. He therefore expands his candidate belief attributions accordingly. He can then note other actions which the subject should perform given the beliefs and desires attributed to him, and test his attributions by observing the creature further or perhaps interacting with the creature to put it in situations which, according to his working theory, should elicit particular actions. The theory can then be improved through a process of reflective equilibrium. It seems that SRC does provide quite a bit of guidance in interpretation.

On this picture, it is supposed to make sense that thinkers can make occasional mistakes, either in practical or theoretical reasoning: a few broken rules do not disrupt the pattern too much, and so we can discover and make sense of thoughts which do have particular contents but which are not related to each other or to action in quite the right way. Too much rule-breaking, however, destroys the pattern, and that is why thinkers are required to approximate perfect rationality.

Again, there is more than one way to challenge the picture at this point. One option would be to point out the fact that we ordinary interpreters and thinkers do not have complete theories of reasoning, and therefore do not have a complete account of the

relevant patterns. This issue is considered briefly in Heal (2003: 229). Let us set it aside for the moment, and assume that SRC, if true, can fulfil the role that interpretationism requires of it.

Another way to evaluate this picture is to consider the extent to which we humans instantiate the patterns, given what we already know about them. If we don't and/or can't instantiate the patterns, we have a choice between saying that we don't count as thinkers, and saying that this picture requires the wrong things from thought. This strategy is taken up in the next section.

3. Normal human irrationality

There are many things within the range of normal human thought that might be called irrational. In this section I concentrate on two of them: the two that are most relevant to challenging a Rationality Claim which uses the Standard Picture.

3.1 Human limitations

Various philosophers have commented on,[76] and many more must have noticed, the limitations of human thought. We hold inconsistent sets of beliefs, often without realising it. And even if we tried to root out all the inconsistencies in our thought, we probably couldn't succeed. Moreover, we definitely do not believe all the logical consequences of our beliefs, and are certainly incapable of doing so, due to our limited powers of deduction. Yet logical consistency and deductive closure are two frequently suggested requirements of rationality.

This offers a first suggestion that we humans might be incapable of doing what rationality demands. Other phenomena also suggest that we sometimes do what rationality forbids.

[76] See Cherniak (1986), Stein (1996) and Heal (2003).

3.2 Data from the heuristics and biases program

In the latter half of the Twentieth Century, psychologists began to publish findings which were intended to suggest that people reason in ways that violate the demands of rationality. A research tradition, often called the 'heuristics and biases program', developed, and there is now an extensive literature on the subject. I will outline two of these studies as examples, before discussing their imputed significance.

First, we may consider conjunction problems. These are supposed to test a subject's probabilistic reasoning, and count as tests of theoretical reasoning. Here is one such task presented to subjects by Kahneman and Tversky, and recorded in their (1983):

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.
>
> Linda is a teacher working in an elementary school.
> Linda works in a bookstore and takes yoga classes
> Linda is active in the feminist movement
> Linda is a psychiatric social worker
> Linda is a member of the League of Women Voters
> Linda is a bank teller
> Linda is an insurance salesperson
> Linda is a bank teller and is active in the feminist movement[77]

[77] Different presentations of this problem in the literature list different options, and indeed Kahneman (2011) says that various versions of the problem were offered in their experiments. In one version, the description was given followed by 'Which alternative is more probable? Linda is a bank teller. Linda is a bank teller and is active in the feminist movement.' (2011: 158) In most groups, this produced the same result as the longer test.

The participants in the study were asked to read the description and then rank the likelihood of the options. The result was that the majority of people rated 'Linda is a bank teller and is active in the feminist movement' as more likely than 'Linda is a bank teller', despite the fact that Linda cannot be a feminist bank teller unless she is a bank teller. Judging that a compound state of affairs is more likely than one of the components of the compound has been labelled the conjunction fallacy.

Next, we may consider so-called preference reversal tests. These aim to show that people's preferences are sometimes inconsistent with each other. Since preference reversal bears on whether the choices for action that people make are consistent with each other, it may be seen as a test of practical rationality. An account of the following test can be found in Kahneman (2011):

> Imagine that the United States is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows:
>
> Version 1: If program A is adopted, 200 people will be saved.
> If program B is adopted, there is a one-third probability that 600 people will be saved and a two-thirds probability that no people will be saved.
>
> Version 2: If program A' is adopted, 400 people will die.
> If program B' is adopted, there is a one-third probability that nobody will die and a two-thirds probability that 600 people will die.[78]

[78] Kahneman (2011: 368).

When presented with version 1, the majority of people choose program A. However, when presented with version 2, the majority choose program B'. Yet, A and A' have the same consequences, as do B and B'. Thus the way the very same options are presented appears to bring about preference reversal. We can ask which consequences respondents prefer, and thus see that they prefer both (or neither).

Numerous other experiments have been carried out.[79] However, the examples above will be sufficient for discussing the problems posed to the Rationality Claim by apparent systematic irrationalities.

Many people have taken results like the above to support a highly pessimistic view of human rationality, or even to prove that humans are not rational. For example, a newspaper article about Kahneman (2011) contains the following claims:

> [T]he discovery of loss aversion proved to be an important refutation of human rationality. (Lehrer (2012))

> Kahneman and Tversky demonstrated that real people don't deal with uncertainty by carefully evaluating all of the relevant information. They stink at statistics and rarely maximize utility. Instead, their choices depend on a long list of mental short cuts and intemperate emotions, which often lead them to pick the wrong options. (Ibid.)

> Although we'd always seen ourselves as rational creatures—this was our Promethean gift—it turns out that human reason is rather feeble, easily overwhelmed by ancient instincts and lazy biases. The mind is a deeply flawed machine. (Ibid.)

[79] See Kahneman (2011) for further examples.

Psychologists and philosophers have sometimes (but by no means always) expressed their reactions to the results in more moderate terms. Nevertheless, many take the results of the heuristics and biases program to show that we do not even approximate rationality as it is described by the Standard Picture.[80]

3.3 The resulting argument

These common forms of human irrationality, if such they be, pose a problem for the interpretationist, since they can be used to attack the Rationality Claim using something like the following argument:

1. We are not rational;
2. We have thoughts;
3. Therefore having thoughts cannot require rationality.

An interpretationist must reject this argument. To attack the second premise of the argument would lead us to eliminativism. The interpretationist must therefore attack premise 1. One way to do this involves admitting that we are not rational in the way required by SRC, but then rejecting the Standard Picture as the right conception of rationality to use in the Rationality Claim. This strategy is explored in subsequent chapters. For the remainder of this chapter, I consider how a supporter of SRC might reject premise 1.

4. Responses

4.1 SCR and the heuristics and biases program

[80] See for example Stein (1996), Samuels et al (2002) and Samuels et al (2004).

The results of the heuristics and biases program have been considered particularly worrying for optimistic claims about human rationality. In this section, I first consider why they have been considered so important and problematic, and how this impacts on the Standard Picture's ability to perform the task the interpretationist requires of it. I then consider three responses to the experiments: rejecting the principles of reasoning that the participants in the experiments violate; attacking the experiments (which leads to a discussion of the relevance of empirical results for the interpretationist's Rationality Claim); and utilising the 'two system' conception of the mind currently popular in cognitive psychology. I conclude that, although they do represent a challenge to the SRC, it is not as significant a challenge as is sometimes supposed.

The results from the heuristics and biases program purport to tell us about mistakes that most humans make in reasoning. However, everyone knew that people made quite a lot of mistakes before seeing the results. One of the things that was special about these experiments was that they showed that *a lot of people* seem to make *systematic errors*, involving violating quite *simple, familiar rules*. However, this still doesn't fully explain the significance of the results. Compare this with a case where, while investigating the claim that people have mostly true beliefs, we discover that people systematically believe their children to be cleverer than they are. Such a result clearly would not do very much to convince us of the falsity of that claim.

There are two important differences which explain the concern over the results in section 3.2. First, the experiments above may be taken to show not just that we reason incorrectly when presented with these particular (and often artificial) puzzles. Rather, they may be taken to indicate that we are bad at applying certain simple rules that are important *throughout* our epistemic endeavours. Thus, they may show that we probably get a large proportion of our reasoning wrong. Second, although the experiments above cannot by themselves entail very much about human reasoning abilities in general, they may be thought to be best explained by,

and so lend support to, a theory which takes a very pessimistic view of such abilities. I will outline this theory now.

The view of human reason often associated with the heuristics and biases program is that 'ordinary people lack the underlying rational competence to handle a wide array of reasoning tasks, and thus… they must exploit a collection of simple heuristics which make them prone to seriously counter-normative patterns of reasoning or biases.' (Samuels et al (2004: 2)) Another way this is often put is to say that instead of thinking as rationality demands, people use 'fast and dirty' heuristics.

'Rational competence' in the above quote makes reference to a distinction between performance errors and competence errors, mentioned in section 1. The former involve a subject in some sense knowing or being able to use the correct rules for reasoning, but for some reason failing to utilise them (for example because they are not paying attention, or are not motivated to do so, or are drunk). The latter, on the other hand, involve in some sense possessing the wrong rules for reasoning, and getting things wrong for that reason. Its use in this context allows us to suggest the possibility that the sorts of errors we were already aware of gave us reason to think that people made performance errors, but that the systematic way in which people fail the reasoning tests above shows that they are also subject to competence errors. Thus, the idea that we are even capable of being rational is threatened. While performance errors might threaten only the view that we succeed in being rational conceived of as an achievement, competence errors are supposed to threaten also the idea that we are rational in the sense of having a capacity. Understood in this way, the results of the heuristics and biases program directly threaten a Rationality Claim which uses Stein's development of the Standard Picture (which says that approximating the perfectly rational being may involve possessing the same capacities, even if one does not always use them).

Philosophers and others, then, have taken the results of the heuristics and biases program to support the pessimistic view of human reason. There are several ways to respond, and I consider three of them below.

The first approach involves rejecting the principles that participants violate as the right principles to follow. This sort of stance has some merits, as we will see in subsequent chapters. However, in combination with the Standard Picture of rationality, it amounts to the claim that given how humans actually reason, we need to revise our best formal theories.

The problem with this response is that, in the brazen form under consideration, it conflicts very strongly with our intuitions about what counts as good reasoning. Most of us can be brought to see the usual answers given in the experiments as wrong. When presented with the rules of our formal theories, we usually feel quite certain that they are right. We can also see, or be brought to see, how violating the rules sometimes leads people into trouble. Our attempt to develop the Rationality Claim needs to pay some attention to our beliefs about what counts as good reasoning.

The second strategy involves attacking the experiments that were used in the program, or at least trying to reinterpret their data. This is often pursued in a piecemeal fashion, with different experiments being challenged in different ways.

In the case of the conjunction test, one popular suggestion is that the wording of the task generates unintentional conversational implicatures, and thus that the question the participant thinks he or she is being asked actually does have the answer they supply. In the example above, it is suggested that people suppose that 'Linda is a bank clerk' is supposed to imply that Linda is not a feminist. It is also suggested that such implicatures would be very difficult for experimenters to cancel out.

In the case of the preference consistency test, the best strategy would also probably be to suggest that people might be misunderstanding the question: for example, we might suggest that in version 1 participants suppose the consequences of program A to be '200 people will definitely be saved, although more might be

saved'; while in version 2 they take the consequences of A' to be '400 people will definitely die, but more might die.' If interpreted in this way, the choices presented by the two versions would no longer be identical.[81]

The essential point is that each of the experiments can be challenged or reinterpreted so that they do not threaten SRC as seriously. Of course, the challenges and reinterpretations can also be challenged, and further argument generated, which I have not pursued here. Here, I will just note that not all of the challenges need to work: the supporter of SRC can accept that humans make a few mistakes and even lack some few competences. They just need to argue that these are rare enough that we can move towards a more optimistic view of human reasoning, and claim that we are rational in the sense needed for SRC to do its work. The possibility of challenging the experiments, then, at least makes the pessimistic view of human reason look rather less well-supported.

However, as well as specific debates about the individual experiments, there are some general worries about this sort of strategy for protecting SRC. One such worry, offered by Samuels et al (2004) is that if we cannot rely on participants understanding their tasks in the same way as the experimenters, then empirical evidence which appears to support the idea that humans are rational will be undermined as well as the evidence against. Stein (1996), on the other hand, objects to all the replies on the grounds that, since they are different for each experiment, they must be ad hoc and therefore suspect:

> In general, experimental evidence for the irrationality thesis can always be resisted in some such way. Prima facie, however, these resistance techniques offered by friends of the rationality thesis[82] are ad hoc immunization strategies. Without some special theoretical considerations that suggest

[81] Admittedly, this does seem a little desperate.

[82] Stein's rationality thesis is different from what I call the Rationality Claim. The latter says that any thinker must be rational. Stein's rationality thesis, in contrast, says that *humans* are mostly rational. His irrationality thesis is then linked to the pessimistic view of human reasoning.

> otherwise, it seems unlikely that interfering factors are always behind divergences from the norms. (110)

The interpretationist will respond to such challenges by claiming that there *is* a general theoretical justification for trying to interpret a person as rational despite apparently contrary evidence: namely, his preferred a priori argument for the Rationality Claim. We might have to use different strategies to discredit different sets of claims, but this will only be because the different claimants are making different mistakes. We won't have to gerrymander our theory to meet the different challenges, but will reply to each using the same underlying theoretical justification that no empirical data could prove that a group of thinkers were not reasonably rational.[83] Moreover, since there is supposed to be an a priori justification of the Rationality Claim, we do not need to worry too much about discrediting evidence against the Rationality Claim only at the cost of also losing the evidence for it.

I think that there is some mileage in this interpretationist reply. However, it would be a mistake for the interpretationist to claim that empirical data about how we reason has no relevance at all to the plausibility of the Rationality Claim. Interpretationism is supposed to be grounded, at least to some extent, in our everyday interpretations of ourselves and others. One of its attractions, discussed in chapter 1, was that it could provide an account on which we come out as knowing at least some things about our own and other people's mental states; on which we do at least usually think and talk about the things we think we do; on which we sometimes succeed in communicating with each other, etc. Enough empirical results could threaten this motivation for interpretationism. For example, if it were too often the case that the intuitive interpretation of someone's actions made them massively irrational, and the available reinterpretations which could show them to

[83] Stein is quite aware of this reply, and addresses arguments for his rationality thesis that are closely related to the arguments for the Rationality Claim presented in the previous chapter. He finds all of them wanting. The interpretationist, of course, must disagree.

be rational portrayed them as failing to communicate with anyone around them and failing to think about or desire the things they had most contact with, then we would have reason to question our version of interpretationism. When we reinterpret to retain the rationality of our subject, this must not make things too complicated, or counterintuitive. If it does, then we will be rightly tempted to decide either that some people (maybe including ourselves) turn out not to have thoughts; or that the Standard Picture must have been the wrong conception of rationality for the interpretationist to adopt, or that there must be something wrong with the arguments for the Rationality Claim.

Thus, there is something in both Stich and Stein's complaints: interpretationism *does* require empirical support in the form of our experience matching what it has to say reasonably well. Too much evidence which intuitively seems to show that we make massive and systematic mistakes could start to undermine our theory. The data from the heuristics and biases program is therefore relevant to deciding whether or not we should accept SRC. The question is whether there is enough evidence yet to make this line of reply untenable. I will not pursue this question here, since I think that section 4.2 will provide us with enough reason to abandon SRC in any case.

The final response to the results of the heuristics and biases program accepts that the experiments show that people do fail to reason in accordance with normatively significant rules. However, it rejects the idea that this shows that humans lack the requisite capacity. Interestingly, this is the view that Kahneman adopts, despite the more extreme responses his research has prompted. To explain this response, I will introduce the terminology of System 1 and System 2, originated by the psychologists Stanovich and West and now widely used in the field. The terminology is used somewhat differently by different researchers, but I follow Kahneman's usage.

According to Kahneman, 'System 1 operates automatically and quickly, with little or no effort and no sense of voluntary control' (2011: 20). System 2, on the other

hand, 'allocates attention to the effortful mental activities that demand it, including complex calculations. The operations of System 2 are often associated with the subjective experience of agency, choice, and concentration.' (21) Kahneman says that we should see these two systems as 'useful fictions', and that they are not 'systems in the standard sense of entities with interacting aspects or parts.' (29) Rather, this terminology allows us to speak more easily of the two *modes or styles* of thinking that humans turn out to use. Thus, 'System 2 is responsible for calculation' is shorthand for "Mental arithmetic is a voluntary activity that requires effort, should not be performed while making a left turn, and is associated with dilated pupils and an accelerated heart rate." (29) 'System 1 is good at association', on the other hand, implies something like "Connections between ideas often occur to a person quickly and automatically, and without noticeable effort." Kahneman then goes on to use this terminology to explain the results of the experiments he and others have performed.

On Kahneman's understanding, the results of the heuristics and biases program do not show that ordinary people lack the capacity to follow rules of logic, etc., but that System 1 lacks this capacity. He thinks that we often rely on System 1, and so get the wrong answers to certain questions, but that we could engage our System 2, which does have the capacity to follow the rules of the Standard Picture, and thus get the right answers. Thus, if we agree that the interpretationist can make do with a notion of rationality as a capacity, we seem to be able to accept the experimental results and endorse SCR.

There are, however, a few snags with using Kahneman's approach to protect SRC. First, according to Kahneman, because the operations of System 1 are automatic, it offers and so inclines us towards answers that are sometimes wrong. This leads to cognitive illusions which may persist even when we engage our System 2, and which may be very difficult to overcome. A second problem is that Kahneman suggests that we are bad at knowing which sort of thinking we have used to arrive at a given conclusion, making it more difficult for us to assess our own

reliability in a given case, and the third and most serious problem is that System 2 is much slower than System 1, and it takes far more effort to solve problems using it: it could not be used to solve all the problems that we face. The two system interpretation of the results therefore adds to the inventory of human limitations, and these limitations are themselves a serious problem.

4.2 SRC and human limitations

The results of the heuristics and biases program present a challenge to SCR, but not, it seems, a conclusive refutation. The more serious problem, I will argue, arises from our human limitations.

Consistency and completeness frequently appear in discussions of what rationality involves. For example, they are mentioned by both Davidson and Dennett in their discussions of rationality, and in explaining the Standard Picture Heal asks us to

> imagine that we have some package of demands of rationality, including at least consistency and completeness in belief together with some other suitable elements, which is agreed to spell out (at least a substantial part of) what rationality demands. (2003: 230)

Consistency and completeness, then, are definitely among the demands of rationality according to the Standard Picture, and so the perfectly rational being has no inconsistent beliefs, and believes the logical consequences of all its beliefs. The perfectly rational being may make mistakes or remain ignorant about some aspects of the external world, but such ignorance and error is only about the contingent. We, however, do not have perfectly consistent belief sets, and we definitely do not believe all the consequences of our beliefs. Nor are we *ever* going to fulfil these requirements. This truth is so obvious and commonplace that the interpretationist

cannot plausibly dispute it using his preferred argument for the rationality claim to reinterpret what is going on. This is the problem of section 3.1: we humans do not seem to do what rationality demands.

Section 1 characterised the Standard Picture in such a way that it did not demand that we be perfectly rational beings in order to be called rational. Rather, it demanded that we approximate such a being. The fact that we don't instantiate perfect consistency and completeness is not, in itself, an insurmountable problem. SCR could still be true if we ordinary thinkers just believed *most* of the consequences of our beliefs *most* of the time, and hardly ever believed inconsistencies (this was the first suggested understanding of approximation in section 1). Ordinary thinkers do not achieve this much, however: clearly, we fail to believe *most* of the consequences of our beliefs. SCR could also be true if we could say that we have the capacity to eliminate all inconsistencies and to believe all the consequences of our beliefs, but don't use this capacity (this was Stein's suggested understanding of approximation). But this isn't plausible either.

Maybe the proponent of SRC could say that the Standard Picture can allow that even the perfectly rational being might not fulfil these requirements until it reaches the end state mentioned above. Unlike the perfectly rational being, we may never actually reach the end state, but as long as we are making good progress towards consistency and completeness, and would achieve it eventually, given enough time, then the Standard Picture will be able to count us as rational. The problem is that the ordinary human does not even seem to be *on the path* towards perfect consistency and completeness in belief. No matter how long a human lives, they will not even nearly achieve consistency and completeness in their beliefs. They will not deduce even half of the consequences of their beliefs, and a long life actually seems likely to increase the number of inconsistent pairs of beliefs.

Given their prominence in discussions of rationality, one might take completeness and consistency to be two of the most important demands of rationality. Yet, we do not have even the capacity to fulfil them if our lives are

radically extended. This shows that we are not rational in the sense given by the Standard Picture, and so that we should not accept SRC.

Moreover, it is not only the case that we are not on a path to consistency and completeness: an ordinary human would not even want to be on that path. Even trying to maximise the degree to which we meet these requirements would take up all of our time and prevent us from pursuing other things that are valuable to us. There is a strong temptation to say that it would be stupid for us to try to go down this path: that pursuing perfect consistency and completeness isn't (overall) rational for us. This threatens another claim that Stein included as part of the Standard Picture: that the rules from our best theories of reasoning are all ones that we ought to follow.

As well as admitting that we can be ignorant and mistaken about contingent fact, it seems that any reasonably intuitive account of what is rational for us ought to take into account our limited cognitive abilities. The Standard Picture can allow for our human limitations to some degree: it can accept that we will not ever achieve consistency and completeness. However, it does not seem able to cope with the idea that our limitations should actually affect what rules we ought to try and obey.

5. Conclusion

The Standard Picture offers an initially intuitive way to understand the interpretationist's Rationality Claim, and it appears that it is possible to combine it with an account of how thought content is determined. However, given this understanding of rationality, there are at least two groups of apparent counterexamples to the Rationality Claim. The first involves the evidence from the heuristics and biases program, and purports to show that ordinary humans make frequent and systematic errors in reasoning. I have argued that this set of counterexamples is not as serious as sometimes supposed. However, the second

group, containing the everyday evidence we have for our human limitations, presents a more serious challenge. These show that we are not rational in the way suggested by the Standard Picture, and should not even want to be. Simply pointing to the rules in our best formal theories to indicate what perfect rationality involves, and saying that thinkers approximate this perfect rationality, does not provide the notion of rationality that the interpretationist is after.

For the limited creature, it seems very plausible that good thinking involves the employment of *appropriate* reasoning strategies given their situation and resources. The simple Standard Picture presented at the beginning of this chapter does not have the resources to accommodate this. But maybe this would be a helpful thing for the interpretationist to take into account when developing his Rationality Claim and his account of interpretation. The next two chapters consider theories of rationality which would allow the interpretationist to use this idea.

In the next chapter, I consider a theory which could keep the Standard Picture's commitment to the idea that rational beings approximate perfect rationality, while dropping the claim that rules are important. I argue that this does not provide what the interpretationist needs either.

Chapter 9 - Consequentialism

The consequentialist about rationality says that being rational is a matter of thinking and acting in ways that achieve or have a good chance of achieving certain valuable ends. A Rationality Claim which uses this conception says that all thinkers must think and act so as to achieve or have a good chance of achieving the relevant goals at least reasonably well given their resources. I will call this consequentialist rationality claim CRC.

One of the advantages claimed for the consequentialist conception is that it meets what Samuels et al. call the Value Condition: 'A normative theory of reasoning should provide us with a vindication of rationality. It should explain why reasoning in a normatively correct fashion matters – why good reasoning is desirable.' (2004: 40) They believe that this condition tells against *any* deontological view, because where rules lead to good consequences, the consequences will be what justify a subject in following the rules, while whenever following a set of rules fails to yield good consequences, it will be unclear why following the rules should be valuable. Thus, consequences always take priority over rules when we decide what to do or think.

There are things that can be said in response to this argument on behalf of the deontologist, and some of them will be said in the course of this chapter. In particular, I argue that if you want a notion of rationality to do a certain theoretical job (namely, to determine the nature and content of thought), then rationality cannot amount merely to the tendency towards the attainment of consequences.

Section 1 begins by showing that there are several importantly different ways of filling out a consequentialist conception of rationality, and section 2 shows how this collection of notions could each be used to deal with the results of the heuristics and

biases program and facts about human limitations without giving up the Rationality Claim. So far, so good for CRC. However, section 3 then argues that purely consequentialist notions of rationality cannot do the work that interpretationism requires of them.

1. Some varieties of consequentialism

The most obvious point of potential disagreement when we come to fill out a consequentialist conception of rationality is what consequences rational thoughts and actions are supposed to constitutively lead towards[84]. In ethics, there are several forms of consequentialism because there are several different accounts of what consequences constitute 'the good'.[85] The situation is the same here: people may have different ideas about what count as good consequences of thinking and acting.

A first sensible suggestion is that rational thought might constitutively lead towards some epistemically valuable consequences, such as the acquisition of true beliefs. However, this seems to have the potential to cover only the goals of theoretical rationality.

An alternative suggestion is that rational thought and action constitutively lead towards those things which we already value, whatever they may be. A consequentialism which employs this suggestion seems particularly well suited to satisfying the Value Condition. It also reduces the universality of the demands of rationality: since people value different things, their being rational will involve different things. This is one of the ways in which, for a consequentialist, rationality can be relative to a subject. One representative of this option in the literature is Kacelnik's (2006) notion of E-rationality (economic rationality), which he defines as the maximisation of expected utility. A similar idea also seems to be suggested by at

[84] What this 'leading towards' amounts to may also be interpreted differently by different consequentialisms. It is deliberately left vague here, but briefly discussed later in this section.
[85] For some options, see Parfit's 'What Makes Someone's Life Go Best' in his (1984).

least some comments made by Dennett, for example in one of his explanations of the relationship between his intentional and design stances. Dennett says that when using the design stance, 'one ignores the actual (possibly messy) details of the physical constitution of an object, and, on the assumption that it has a certain design, predicts that it will behave as it is designed to behave under various circumstances' (1987: 16) He then says both that the intentional stance involves treating an object as a rational agent, and that 'One can view the intentional stance as a limiting case of the design stance: one predicts by taking just one assumption about the design of the system in question: whatever the design is, it is optimal.' (73) If Dennett then attributes desires for the consequences (likely to be) achieved to subjects, it seems that for Dennett being rational amounts to achieving those things which you are optimally designed to achieve, which therefore count as the things you desire.[86]

A third option would be, rather than focusing on what we actually do value, to refer to the 'ends' of evolutionary processes such as fitness and the survival and perpetuation of genes. Such a notion is also developed by Kacelnik (2006). He calls this B-rationality (biological rationality), and explains how individual creatures can be seen as maximising agents for the alleles they carry. A similar notion of rationality also seems to be at work in the optimistic view of human rationality presented by some in the evolutionary psychology research program.[87]

Finally, we could say that rational actions constitutively lead towards the promotion of some sort of 'objective well-being' of the subject, where well-being is explained in terms of more than the 'goals' of evolutionary processes and/or the actual desires of the subject.

It is important to recognise these different options, because different things count as rational for a creature depending on which of them we choose. For

---

[86] This is only a brief sketch of something Dennett *might* be taken to think. More would need to be done to develop the above into a plausible picture, or to argue that Dennett is committed to this. Although it fits with some of Dennett's comments, it should also be noted that it goes against others, such as his claim that he depends on a 'systematically pre-theoretical' (1987:98) notion of rationality.
[87] Some discussion of which can be found in Samuels et al (2004).

example, imagine an environment containing two kinds of snakes which are very similar in appearance, but only one of which is poisonous. Then consider two people in this environment, one of whom infers that he is in danger whenever he sees a snake, and runs away; and the other of whom will not infer that he is in danger until he has spotted a distinguishing mark on the underbelly of the snake. If reasoning and action which lead to the avoidance of false beliefs are rational, then the second man is the rational one. If reasoning and action which leads to the satisfaction of a subject's desires are rational, then either (or both, or neither) may be rational, depending on what each desires. If reasoning and action which leads to the promotion of the ends of evolutionary processes or the objective well-being of the subject are rational, then the first is surely more rational than the second.

Another important question is that of the environment relative to which rationality is to be determined. We could say that the degree of rationality must be always be assessed relative to the same environment; our environment perhaps. Or, we might try to assess the rationality of each creature by looking at all situations in which it would be logically possible for them to exist. More intuitively plausible options involve making the appropriate environment relative to the subject whose rationality is being assessed, thus introducing another way in which the demands of rationality may be relative. For example, we might assess a creature's rationality relative to its actual environment: i.e. determine whether its reasoning and actions led towards the chosen positive consequences in whatever situation they were actually in. Something like this option appears to be employed by Bennett (1976).[88] Another option would be to assess rationality relative to the environment in which the creature in question grew up or has spent most of its time, and another is to say that rationality is determined relative to the environment in which a creature's

[88] Bennett here isn't trying to provide a theory of rationality, which he thinks involves language as in his (1964), but he is providing a theory of teleology which he hopes will serve as a theory of thought. I think that he is using something that could quite plausibly be called a notion of rationality, but I would not count it as a purely consequentialist one.

ancestors evolved. Whichever the consequentialist chooses, they then need some way to delineate environments.

Once again, different things will count as rational for a creature depending on which environment we take to be relevant. Contrast the previous example of the people and their responses to snakes with the following: imagine an environment containing two types of non-poisonous snakes. If we assess rationality relative to current environment, a person in this new environment who, upon seeing a snake, infers that they are in danger and runs away, may well not do what best promotes his evolutionary potential, since he may expend energy unnecessarily and end up avoiding places which could be useful to him.

Finally, we may also ask how we are to interpret the vague term 'lead towards' when speaking of whether thoughts and actions lead towards certain consequences in particular environments. One issue here is whether the thoughts and actions must actually or necessarily result in the consequences, or whether they must raise the probability of the consequences by some degree. If the latter, we can then ask whether we are here using an objective notion of probability, or a subjective notion, and if a subjective notion, from whose point of view is it to be measured? This introduces a question about the importance of the point of view of the subject on the consequentialist view, which I will not attempt to answer here.

Each of these proposals can be criticised in various ways as analyses of our everyday notion of rationality, and even within the consequentialist options, there is no need to argue for one real or somehow superior notion of rationality. Rather, we may use different notions for different purposes, as Kacelnik (2006) does. However, once we have a purpose in mind, as the interpretationist does, it is important to specify which notion we are using just because the various possibilities label different things as rational. As I will also argue in section 3, different versions of consequentialism fail the interpretationist in slightly different, though related, ways. Nevertheless, the

various consequentialist proposals do have enough in common to treat their responses to human limitations and the results of the heuristics and biases program together.

2. CRC, human limitations and the results of the heuristics and biases program

CRC says that all thinkers think and act so as to achieve or have a good chance of achieving or at least increase their chances of achieving the relevant goals at least reasonably well given their resources and an appropriate environment.

Unlike SRC, this does not face a problem in accepting and accommodating our human weaknesses. For example, the fact that we do not achieve consistency or deductive closure in our beliefs does nothing to disprove CRC, because a consequentialist notion of rationality takes what we are capable of doing into account when determining what is rational for us. In other words, a consequentialist picture seems to be able to do exactly what I complained that the Standard Picture could not do in section 4.2 of the previous chapter.

Given CRC, it is also far more difficult to say that the experiments detailed in chapter 8 prove that we suffer from counter-normative reasoning competences. As Samuels et al. say,

> We are no longer in a position to confidently invoke familiar formal principles as benchmarks of good reasoning. Instead we must address a complex fabric of broadly conceptual and empirical issues in order to determine what the relevant standards are relative to which the quality of our reasoning should be evaluated. (2004: 45)

Whether the reasoning strategies employed by participants in the experiments above should count as irrational comes to depend on their individual cognitive capacities,

on whatever is the correct environment in which to determine the likely success rate of their strategies, and on what consequences their strategies are supposed to achieve. This makes it possible that the so called 'fast and dirty' heuristics identified by the heuristics and biases program are in fact the most suitable rules for us to use, and so a sign of our rationality, rather than a disproof of it. A more detailed response to the results depends on one's chosen form of consequentialism.

As Samuels et al. say this form of response does not make empirical results such as those from the heuristics and biases program completely irrelevant. Given any particular understanding of the goals of rational thinking and our limitations, it may look as if we do not do the best that we can do. At this point, the supporter of CRC may need to start employing some of the manoeuvres I suggested for the supporter of SRC. The consequentialist may be able to employ the second two responses more successfully: if they choose to challenge the interpretation of the results, they can introduce the idea that such results must influence what we take our resources and limitations to be, while if they adopt the third option, they will have less difficulty in accepting the human limitations it implies.

Thus, if we use a consequentialist notion of rationality we seem to be in a reasonably strong position to deal with various empirical facts about humans and their reasoning capacities. The next question is whether the consequentialist conception can do the work that interpretationism requires.

3. Rejecting CRC

To recap, the consequentialist conception of rationality says that being rational involves thinking and acting in ways that lead towards certain goals given the available resources. On this account, being more rational involves getting closer to doing and thinking whatever would best achieve the specified goals in the relevant environment; while being less rational involves not doing the best thing all the time,

and perhaps sometimes doing completely the wrong thing. CRC says all thinkers must think and act so as to achieve the relevant goals at least reasonably well given their resources.

We have seen two reasons why the interpretationist might favour CRC over SRC: a consequentialist notion of rationality seems better able to meet the Value Condition, and CRC is better able to take into account our human weaknesses. Here, I show that CRC has a very major disadvantage; it does not allow the interpretationist to give an account of how the content of thought is determined.

In chapter 7, section 2.2 we considered an argument for the Rationality Claim which called on the requirements of interpretation. This argument is of the utmost importance here, since it tells us how the interpretationist wants to use the Rationality Claim. Let us begin by rehearsing it. The argument started from the assumption that thinkers had to be interpretable, and then pointed out that interpretation requires known connections between a creature's environment, their actions and what they think. It suggested that the connections were provided by rationality, and so concluded that thinkers must be rational. What the interpretationist needs to show is that a consequentialist notion of rationality gives us what we need in order to produce interpretations.

Consider again the Analogoids from the previous chapter. They are octopus-like creatures who behave as follows: most of the time, they move slowly through the water and ingest tiny creatures near the ocean bed, or float near the surface of the water in the sunshine. Except, that is, for when one of three predators enters the vicinity. Analogoids move upwards when approached by a giant sea spider, which is fast but can only move along the ocean floor. They swim quickly away from approaching sluggish eels, which can swim anywhere, but move slowly. And they swim in amongst the branches of a plant on the ocean floor when approached by the fast shark-like predator which can swim in all directions, but which is too large to enter the branches of such plants.

Suppose we want to use the intentional stance to interpret an Analogoid. It seems like such an interpretation is very easy to give: we can say that when it sees a giant sea spider and then swims upwards, it does so because it thinks that there is sea spider coming towards it, and that it is therefore in danger, it thinks that swimming upwards will keep it safe, and it wants to remain uneaten; when it sees a sluggish eel and swims quickly in the other direction, this is because it thinks it is in danger from an eel, wants to remain uneaten, and thinks that swimming away quickly will keep it safe; and so on, with the addition of thoughts about the sharks and how to escape them, and various additions about the Analogoid's other desires, how they are ranked, and how it thinks it can best fulfil them. Such is the obvious intentional interpretation of such a creature. However, CRC does not provide us with enough information to produce this interpretation.

If we assume only that thinkers must think and act so as to achieve the relevant goals at least reasonably well given their resources, we are equally entitled to give the following interpretation: when the Analogoid sees a giant sea spider, it believes itself to be safe, desires to remain uneaten, and believes that swimming into the plants will put it in danger, it decides on the basis of this that it ought to swim forwards very quickly, and as a result of this intention swims slowly upwards; when the creature sees a sluggish eel, it forms the belief that there is food nearby, desires to remain uneaten, and believes that swimming round very slowly in circles will be amusing, it infers from this that it ought to bury itself in the sand, as a result forms the intention to swim into the plants, and as a result of this intention, it swims very quickly away from the sluggish eel; and so on.

Now, there's clearly something wrong with this interpretation, and we might explain what is wrong by saying that *obviously*, a rational creature won't form the belief that it is safe on meeting an often-encountered predator, o*bviously* it won't decide on the basis of the belief that it's safe where it is and the desire to remain so, that it needs to swim quickly, and *obviously*, on forming the intention to swim forwards quickly, its intentional action cannot consist of swimming slowly upwards.

The same goes for the interpretation of the Analogoid after it sees the sluggish eel, but perhaps even more so. But the fact that these interpretations are so obviously wrong is not captured by CRC. In the interpretations, the creature has an experience which causes certain mental states, and these mental states then cause it to take the appropriate action to realise the purpose we are attributing as a desire to the creature. This is as much as CRC requires, given the version of consequentialism on which the relevant consequences are the satisfaction of the rational subject's desires. Those versions of consequentialism in which the relevant consequences are the ends of evolutionary processes or the achievement of objective well-being seem to provide even less constraints on interpretation, since it seems they may also allow us to make up strange desires for the Analogoids. What happens when we emphasise epistemic consequences is dealt with below.

The basic problem that we have encountered here is that consequentialism, due to its focus on outcomes, seems unable to say anything about how the thoughts and actions which lead to an outcome ought to be related to each other. And this means that, using only CRC, the interpreter can't determine what the thoughts are.

At this point, perhaps a supporter of CRC will say that a Rationality Claim doesn't need to provide *all* that an interpreter needs in order to interpret their subjects. Perhaps CRC is essential to interpretation, but other assumptions, having nothing to do with rationality, are also needed, and will be able to resolve the indeterminacy encountered above. However, the problem we are encountering is that CRC does not inform us how thoughts have to relate to each other, and this seems to be exactly the sort of thing that a Rationality Claim *should* tell us. Indeed, it is something about which other Rationality Claims have a lot to say. I therefore reject this escape route from the current problem.

Alternatively, the consequentialist might object to the argument by saying that the problem arises not from using a consequentialist notion of rationality, but from the kind of example to which it is applied. In particular, they might say that the problem arises because we are considering creatures which do not have a

language. We cannot choose between rival interpretations of them, but that is just a familiar problem with the interpretation of non-linguistic creatures, and may simply show that we should, with Davidson, deny that such creatures really possess thoughts at all.

However, the indeterminacy here is much more extensive than that identified by Davidson. When considering a dog's supposed belief that a cat has gone up a tree, Davidson says

> That oak tree, as it happens, is the oldest tree in sight. Does the dog think that the cat went up the oldest tree in sight? Or that the cat went up the same tree it went up the last time the dog chased it? It is hard to make sense of such questions. But then it does not seem possible to distinguish between quite different things the dog might be said to believe. (2001: 97)

Still, Davidson seems to admit that, to the extent that we can identify what the dog is thinking about, we know it is thinking about the tree. In the case of interpreting an Analogoid using CRC, we cannot determine whether it is thinking about a sluggish eel, a sea spider, or a shark; or indeed an elephant, a planet or its great grandfather. The problem seems not to be that we have a method of interpretation but lack enough behaviour to apply it to, but that we lack a way of generating an interpretation from behaviour. The addition of language would not help here.

But what about the first version of consequentialism suggested in section 1, according to which at least one of the things rational creatures must aim for (and perhaps achieve) is true beliefs? The deviant interpretations above do not involve the Analogoids succeeding at this. Unfortunately, this does not completely resolve the problem. Although some deviant interpretations are ruled out by this version of CRC, we can suggest many others, where the Analogoids form true and comprehensive beliefs that are, however, at the times that they develop and act on them, completely disconnected from their environment and projects. So, for

example, we could offer the following: when the Analogoid sees a sharkish predator at time t_1, it forms the (true) belief that it encountered a sluggish eel a week ago, and the belief that the best way to escape from giant sea spiders is to swim upwards. It then combines these with the desire to remain uneaten and deduces the second law of thermodynamics. It forms the intention to assassinate the leader of the Women's Institute, and as a result swims into the branches of the nearby plants.

What we need is to be able to demand that beliefs, other thoughts, environments and actions be connected to each other in appropriate ways. But then we must say what these appropriate ways are, and we risk abandoning consequentialism in favour of another account of rationality.

The basic objection that I am presenting here is that rationality, in order to do what the interpretationist requires, must be a matter of the method of achieving as well of the achievement of a consequence. If rationality is connected to thought in the way the interpretationist supposes, then it must be a matter of achieving outcomes in a way that involves states which are related to each other in particular ways. Only then can such states be thoughts. When you ask what the particular ways of being related are, the most obvious answer appears to be 'the rational ways', suggesting that consequentialism does not give us a complete account of the notion of rationality needed by the interpretationist.

To avoid the conclusion that consequentialism does not give a sufficient account of rationality for our purposes, what is needed is an argument that having certain kinds of thoughts with certain connections between them is the best way of achieving any outcomes that might be relevant.

I suggest that at this point the consequentialist might be tempted to point out that he need not eschew the rules of the Standard Picture, or some other set of rules, entirely. If following them is useful, then the creature which satisfies CRC should conform to them when resources allow. The consequentialist may not be able to point to an easily identifiable abstract entity the instantiation of which results in

thought (as did the supporter of SRC). However, he may suggest that this does not mean that on his theory there is no complex pattern that each thinking creature needs to instantiate. This pattern might be the one a creature instantiates when it follows a very complicated set of rules. There might be some set of rules which would say when a creature should and shouldn't obey the rules picked out by the Standard Picture, when it should satisfice in its pursuit of its goals, and so on.

The consequentialist may then argue that admitting the existence of such a set of rules need not make us into deontologists about rationality: we could say that although there were rules, they were the rules that mattered to rationality only because they led towards the relevant consequences. It therefore seems as if consequences could still have priority on this theory.

The picture we require, then, is one on which the consequences, attainment of which we see as constitutive of rationality, also determine that a creature needs to follow certain rules in their thought and action. These rules can then provide us with the connections between thoughts which we need to determine the nature and content of thoughts. Moreover, if there is a set of such rules, and we can attribute (probably implicit) knowledge of them to interpreters, we will have a way to rehabilitate the argument from the requirements of interpretation. The question is whether the consequentialist can provide an argument to say that creatures must follow/conform to some set of rules in order to optimally achieve those consequences (tendency towards the) attainment of which is constitutive of rationality.

If the (tendency towards the) achievement of consequences is all that matters for rationality, why should creatures follow a set of rules to attain these consequences? Achieving them in a different way each time is still achieving them. Perhaps the obvious answer to this is to point to the fact that all the potentially rational creatures we have encountered are finite creatures. And finite creatures, with a limited number of resources available to them, and an inability to see the future, need to have finite procedures which they can follow, which enable them to achieve the important consequences most of the time given their environment.

Indeed, this was precisely the point of the consequentialist response to the data from the heuristics and biases program (given in Section 2): it may be beneficial for a limited creature to adopt certain strategies in thought, even when such strategies are not the most reliable ones available. When we responded to the data from the heuristics and biases program, it was important to stress that the strategies need not always involve valid inferences. Here, on the other hand, we may stress that sometimes following valid inferences does provide a way for a creature to achieve its goals.

In reply to this argument, it is sufficient to point out that if it works at all, it only tells us that the thoughts of a finite creature must follow *some* sort of pattern. It cannot tell us what that pattern might be, beyond saying that the pattern must in fact promote the relevant consequences. For example, as long as it fits into *some* pattern of reasoning, it cannot say what is wrong with our Analogoid reasoning like this:

> when it sees a sluggish eel, it forms the belief that there is food nearby, desires to remain uneaten, and believes that swimming round very slowly in circles will be amusing; it infers from this that it ought to bury itself in the sand, as a result forms the intention to swim into the plants, and as a result of this intention, it swims very quickly away from predator 2.

Yet this hardly improves our position at all. We want to say that there are other facts about how thoughts need to be related to each other, and I maintain that pure consequentialism does not have the resources to do this.

4. Conclusion

In conclusion, a purely consequentialist notion of rationality does not provide the interpretationist with the resources he needs to explain the nature and content of thought. It should be noted that this is not a wholesale objection to such a notion of

rationality: it may be ideally suited to many purposes. Samuels et al., for example, are not interpretationists and it remains possible that their conception of rationality does all that they want it to. There also seems to be something right about consequentialism: it seems plausible that rationality should (at least sometimes) have something to do with what is good for us. However, in order to develop a plausible form of interpretationism, we need more than this.

Chapter 10 – Achievements, Patterns and Purposes

The Standard and Consequentialist Pictures of rationality each have attractions. If we take either of them alone, however, they provide us with unsuitable conceptions of rationality for the interpretationist's purposes. This chapter will attempt to suggest a conception of rationality that is more suitable for interpretationism. As a single chapter, it obviously cannot give a complete account of what the interpretationist notion of rationality involves and requires, but it will try to provide the broad outline of such an account, suggesting how we can combine recognition of the importance of rules with recognition of the extent to which they can be broken, sometimes for good reasons, and giving pride of place to the role that an assumption of rationality is supposed to play in understanding others.

I begin in section 1 by returning to an issue briefly raised in chapter 8: whether the interpretationist should see the rationality of a thinker as an achievement or as a capacity. I argue that the interpretationist must include some achievements in his conception of rationality. Section 2 then suggests an 'achievement' that any thinker must attain by suggesting that the rational being must, at least sometimes, do at least some of the things that the Standard Picture says that rationality involves. Section 3 adds to this that there must also be some pattern to or consistency in the way that a creature reasons. Section 4 then shows that, by including these two elements in our conception of rationality, we can give a good response to the challenge from chapter 8, concerning human limitations and errors in reasoning.

In section 5, I consider whether this picture of minimal rationality includes all that we need in order to see those who instantiate it as thinkers with perspectives and reasons of their own. I consider further additions that we might make to the

picture, but argue that it would be better to refrain from making a decision between these related theories, and to endorse a stratified conception of rationality.

Section 6 completes the chapter by considering the purpose of interpretation, conceived of as an attempt to understand a creature in terms of its reasons through assuming it to be rational in the sense argued for.

1. Rationality as an achievement

The issue of whether rationality should be seen as a capacity or an achievement first appeared in chapter 8. There, we were worried about evidence that we do not always reason in a normatively correct fashion. This was met with the suggestion that perhaps it is enough that we could reason better, even though we often don't. The idea that this might allow us to count humans as rational in the sense required by the interpretationist might be compared to the way, in Part II of this thesis, the potential for us to produce interpretable behaviour was taken to be sufficient for thought, even when no actual behaviour occurred. I come now to the evaluation of this suggestion: are the two cases comparable? And is the potential to be rational all that the interpretationist needs a thinker to possess?

We should first note that the intuitive plausibility of the two claims is very different. In Part II, the suggestion was that one might have a thought which one did not display in any behaviour, but which one could display in behaviour, for example if the circumstances were different. I take it that this will seem quite reasonable to more or less everyone. On the other hand, the suggestion that the interpretationist's kind of rationality might be present *only* as a capacity with respect to a certain thought, such as a belief, amounts to the idea that a subject could have a belief which they have not arrived at by rational means and which they then do not use in any of the ways sanctioned by the rules in the Standard Picture, despite having opportunities, but which they *could* nevertheless perhaps form and certainly use in

an appropriate way. Thus, the belief won't have any of the correct upstream connections to other thoughts, such as being made probable by other beliefs etc., and it won't have the correct downstream connections to other thoughts, such as the subject going on to believe any of the consequences of the belief, or the consequences which follow from combining that belief with others of their beliefs, nor their rejecting any of the other beliefs they have which are incompatible with this belief, nor combining it with their desires to form intentions that are likely to lead to the fulfilment of those desires, if the belief is true. Nevertheless, the belief is supposed to be the belief that it is and to have the potential to have (at least some of) the appropriate relations to other thoughts and to actions. While the idea from Part II seemed obviously sensible, the counterpart involving rationality seems to me to be quite implausible.

The idea that we should allow rationality as mere capacity, then, seems unpromising. But why is it so much less plausible than its counterpart? One might suggest that if you think the mere capacity for rationality is enough for belief, then you must have misunderstood the nature of belief: beliefs just aren't the sort of thing which you can have, but not have formed appropriately nor use to guide your thought and action even when you can and should do so. This, however, appears to restate the problem. Certainly, it does not seem to offer anything useful in responding to a person, if such there be, who finds the idea of a belief which is not actually rationally connected to anything else unobjectionable. What we want is to understand why, if they are the sort of thing a subject can fail to display in their behaviour, beliefs are not also the sort of thing one can just fail to form or employ in a rational way.

I think that the interpretationist can answer this question by considering what is different when an interpreter considers the belief that does not result in action versus when he considers the belief that is neither formed nor used in accordance with rational norms. When the interpreter thinks of a subject who is, in fact, failing to display behaviour which can be interpreted to reveal their mental states, the

interpreter takes the situation to involve the subject being, in whatever sense is required for interpretationism, rational. Such situations involve subjects failing to act because they are paralysed, or producing misleading behaviour because they want to deceive. For such subjects, not acting in a way which makes their beliefs interpretable makes sense, given their purposes, their limitations *and* many of the rules we think that rational thinking follows. We can make sense of someone failing to act in an interpretable way precisely because it fits into our scheme for making sense of and understanding each other. When we take a subject failing to achieve any degree of rationality with respect to a belief, however, this cannot fit into our way of understanding others. Rather, the situation destroys the possibility of the sort of understanding we are looking for.

This is a distinctively interpretationist answer to the problem: it involves taking beliefs and other thoughts to be parts of a scheme[89] that we use to understand and interact with others. Because the scheme arises from/is embedded within our practices of understanding and dealing with each other, at least sometimes the application of the scheme must deliver determinate answers about what others are thinking.[90] Nevertheless, the scheme allows us to make sense of the idea that sometimes thoughts might be present, without there actually being enough evidence to allow interpretation in the real world. Indeed, this idea is sometimes even useful in the process of interacting with others.[91] The scheme cannot, however, allow the presence of a thought such as a belief in the absence of anything which would make it make sense, or allow it to contribute towards our making sense of the person, which is what the suggestion that rationality could amount to just a capacity and no

[89] I say scheme, rather than theory here because I do not mean to align interpretationism with the theory theory. With 'scheme' I intend to put in mind both theories and some kinds of practices.
[90] This is not an argument for the Availability Claim, but a consideration in favour of the idea that at least sometimes, people actually do display behaviour which allows us to interpret their thoughts.
[91] Indeed, cf. chapter 4, section 1 for the stronger claim that the scheme makes this unavoidable.

more in a particular thinker would involve. This cannot form a part of the project of making sense of each other.[92]

An interpretationist, then, should not find the suggestion that the rationality of a thinker could be a mere capacity appealing; this does not fit well with his account of what psychological concepts depend on and the role that they play in our lives. So, according to the interpretationist, the rationality of a thinker involves at least some degree of achievement. The next question is what a thinker needs to achieve.

2. Sometimes getting it right

Cherniak (1986) provides a suggestion about how much a thinker needs to achieve in the rationality department. His first departure from the Standard Picture presented in chapter 8 is to introduce an emphasis on the actions of a creature, and the potential satisfaction of its desires. Thus, Cherniak suggests the following as an ideal general rationality condition:

> 'If A has a particular belief-desire set, A would undertake *all* and only actions that are apparently appropriate.' (7)

He explains the apparent appropriateness of an action in terms of whether, according to A's beliefs, the action would tend to satisfy A's desires (1986: 7). This notion will be considered in more detail shortly. In addition to the general condition, Cherniak also specifies ideal conditions for certain aspects of thinking, for example the ideal consistency condition: 'If A has a particular belief-desire set, then if any inconsistency arose in the belief set, A would eliminate it.' (17) and the ideal

[92] Note the focus on beliefs here. I will consider other states, such as imaginings, in chapter 11.

inference condition: 'If A has a particular belief-desire set, A would make all and only sound inferences from the belief set that are apparently appropriate.' (13)

These conditions are not as demanding as my initial characterisation of the Standard Picture: Cherniak's ideal rationality condition does not require even an ideally rational creature to achieve deductive closure, since inferences which do not lead to or prevent the creature acting in a certain way cannot contribute to the creature's satisfaction of the ideal general rationality condition. Cherniak envisages a creature performing only certain inferences depending on whether these look like they will help the creature to achieve its desires. Nevertheless, Cherniak says that these conditions are still too demanding for us humans. We do not, and moreover we should not try to satisfy these conditions.

Instead of the ideal general rationality condition, Cherniak suggests that a cognitive theory should use the following:

> 'If A has a particular belief-desire set, A would undertake some, but not necessarily all, of those actions that are apparently appropriate.' (9)

Likewise, he suggests replacing the ideal consistency condition with a minimal consistency condition: 'If A has a particular belief-desire set, then if any inconsistencies arose in the belief set, A would sometimes eliminate some of them' (16) and offers a minimal inference condition: 'If an agent has a particular belief-desire set, he would make some, but not necessarily all of the sound inferences from the belief set which are apparently appropriate.' (10) Each of these conditions is to be supplemented by the further requirement that the creature must not make too many mistakes: it must not perform too many inappropriate actions or inferences, nor fail to eradicate too many inconsistencies. Cherniak's basic method for transforming ideal rationality conditions into minimal rationality conditions that we can satisfy is therefore quite simple: it is to replace 'all' with 'some' in statements of these conditions.

Cherniak's proposal could be seen as a very cut down version of the Standard Picture, on which 'approximation' to perfect rationality required only doing *some* of the right things. However, we should note that Cherniak's theory can allow that there might be kinds of good reasoning, or occasions of correctly responding to reasons, which can't be captured in rules or principles at all, yet which still exhibit a creature's rationality - performing these kinds of reasoning and these responses could be counted as things which help a creature fulfil the minimal general rationality condition. This represents a very significant departure from the Standard Picture as presented in chapter 8, as well as an additional advantage of Cherniak's theory.

I will call a rationality claim which uses Cherniak's notion of minimal rationality MRC. Let us consider, then, whether an interpretationist could use MRC to produce interpretations. Below, I first consider how using MRC helps us to avoid some of the deviant interpretations that troubled us when using CRC, and then consider three potential objections: that Cherniak's account cannot provide enough for successful interpretation because (unlike the consequentialist picture) it does not allow us know that the creature is acting in a way which really does tend towards the satisfaction of its desires; that the notion of apparent appropriateness is unclear; and that Cherniak's use of terms like 'some' introduces unacceptable ambiguity into our psychological concepts.

Given the notion of 'apparent appropriateness', being rational has something to do with satisfying desires, just as some of the consequentialists in the previous chapter claimed. But it only involves satisfying them by doing what, according to certain beliefs, will satisfy them. The beliefs in question are those of the subject being interpreted. The notion of apparent appropriateness, then, can help us to rule out some of the deviant interpretations that troubled us when we tried to use a purely consequentialist conception of rationality. For example, I worried that the consequentialist could not rule out the following interpretation of the Analogoid:

> When the creature sees a giant sea spider, it believes itself to be safe, desires to remain uneaten, and believes that swimming into the plants will put it in danger, it decides on the basis of this that it ought to swim forwards very quickly, and as a result of this intention swims slowly upwards.

The addition of apparent appropriateness can help us to rule out this interpretation because it does not include the Analogoid doing what is apparently appropriate given its beliefs, but doing the right thing in spite of beliefs suggesting that other actions would be more appropriate.

Nevertheless, as it stands, Cherniak's conception of minimal rationality certainly does not guarantee the production of sufficiently determinate interpretations. Depending on what a subject believes, one set of movements might count as various different actions.[93] Take the example of the Analogoid moving slowly upwards through the water in the presence of a giant sea spider. This could be an act of escaping from a spider. However, it could instead be an act of trying to escape from either of the other two predators, if the Analogoid is confused about both which kind of predator is present, and which predator can go where. Or it could be an act of trying to attract a mate, if the Analogoid believes there is another Analogoid watching who will find such behaviour appealing. Given the right set of strange beliefs and desires, it could even be an act of trying to arrange clothes in a wardrobe, or of attempting to steal a car.

This is the first potential objection mentioned above. As well as gaining information about how to interpret, it seems like we've lost something which the consequentialist account gave us: the assurance that the creature is probably (or sometimes) doing the *right* thing to satisfy his desires. We have lost this because of Cherniak's focus on creatures doing what is *apparently* appropriate, rather than just

[93] Cf. Anscombe's famous example of the man and the water pump in her (1957).

appropriate. To respond to this problem, we must add to the list of things which a thinker must, at least sometimes, achieve. I will consider two ways in which we might do this.

A first suggestion is that we could say that the creature must, at least sometimes, do what is really appropriate for the satisfaction of its desires, i.e. what really will tend to satisfy its desires. This brings us up against a question which we need to ask with respect to the general rationality condition in any case: do the actions of a creature which are mentioned in such minimal rationality conditions need to be observable by an interpreter? Or might purely mental actions, such as deciding to think about cosmology, count as well? If the latter, then the condition that a creature must, at least sometimes, do what is *actually* appropriate for the satisfaction of its desires, will not help us to overcome the problem in interpreting, since we get no guarantee that any of the actions we can observe are among the genuinely appropriate ones. On the other hand, if we say that 'external' observable actions are required, then observable actions become a condition of rationality, and we stop being able to count our locked-in cosmologist from chapter 5 as rational, and therefore as a thinker. Since we want to count the locked-in cosmologist as a thinker, we should not make genuinely appropriate observable actions a part of what any thinker must achieve. Instead, we need to find a way to ensure that for any thinker who does produce observable behaviour, some of that behaviour will be genuinely appropriate (i.e. a good way of getting them what they desire).

A second suggestion is that we add in conditions which ensure that the minimally rational creature has a fair number of true beliefs on which to act. That any thinker must have mostly true beliefs is something that both Davidson and Dennett maintain. Dennett says that one should 'attribute as beliefs all the truths relevant to the system's interests (or desires) that the system's experience to date has made available' (1987: 18) and that 'the attribution of false belief, *any* false belief, requires a special genealogy, which will be seen to consist in the main in true beliefs.' (Ibid.) Davidson, on the other hand, makes attributing true beliefs a major part of

obeying his Principle of Charity. Adding this sort of feature to Cherniak's account of rationality will involve adding conditions about the 'upstream' connections that thoughts such as beliefs are supposed to have, and saying that a minimally rational creature must sometimes get these things right. Cherniak admits that he does not pay much attention to such upstream connections. He is more interested in the 'forwards looking', or downstream connections of thoughts, a focus which makes sense given that his primary interest is in prediction on the basis of existing knowledge of a creature's thoughts. For the interpretationist, however, the condition that creatures sometimes form the right beliefs given their environment will be very important.

The second potential objection to using MRC also focuses on the use of the notion of apparent appropriateness. The problem is that it is not completely clear what 'apparent appropriateness' amounts to, and there are ways of developing this notion which would lead us into difficulties. As I said above, Cherniak says that an action is apparently appropriate for a creature if according to the creature's beliefs, that action would tend to satisfy the creature's desires. But which of the beliefs of the subject are being called upon here? A first suggestion is that all of them are, but a subject's beliefs might (indeed, are likely to) include some contradictions. According to some logical systems, everything then follows from the subject's belief set; according to others, nothing does. Both are unacceptable options.

A better suggestion is that the notion of apparent appropriateness directs us towards just those beliefs which the subject uses to decide upon their action. However, then the first objection surely returns to haunt us. We can suppose that the Analogoid forms some beliefs as it should, but also forms other beliefs in completely the wrong ways. We can then suppose that when it acts, it may only act on the beliefs it formed in silly ways. And in that case, we don't have any way to work out what the creature is trying to do, or what many of its beliefs and desires are. To address this problem, we must look at how we should understand terms such as 'some' and 'not too many' in Cherniak's minimal rationality conditions.

This brings us to our third potential problem. According to Heal (2003), Cherniak's use of terms like 'some' and 'not too many' is problematic, because we could interpret these terms in different ways. For example, when we say that a thinker must make at least some sound inferences, does this mean that some proportion of the inferences they make need to be sound (half, perhaps, or a quarter)? Or does it mean that they must at least succeed in making the more obvious sound inferences that they could make? If we can give several reasonable interpretations of what counts as 'some' etc. in Cherniak's definitions, or if different interpretations of these terms are appropriate for different situations of interpretation, Heal worries that there will be ambiguity in our concept of rationality. Since the interpretationist hopes to illuminate or define thought by considering rationality, the worry is that our psychological concepts will then also be problematically ambiguous.

It is certainly the case that people do achieve very different amounts with respect to keeping the rules of the Standard Picture, and moreover that we hold different people to different standards and assume different things about what they are capable of when we interpret them. For example, we expect different standards from a professor and a three-year-old child, and this surely influences the way we then understand particular of their utterances and actions. The question is whether, in employing the assumption that they are minimally rational, we then understand 'some' differently for each of them. If so, both will be minimally rational, but what we take them to do to achieve this minimal rationality will be quite different. If this difference is quite radical we may then wonder whether the thoughts we attribute to each, such as beliefs and desires, should really count as being the same sorts of states.

One response to this would be to say that drawing a distinction between what it means for a young child to have a belief and what it means for a professor to have a belief is a good thing: the states do, after all, have important differences between them. Still, we want to retain at least some unity in our psychological concepts. But on the current picture, we have a way to do this: there remains an important sense in which we have unified concepts of rationality and belief etc., since these concepts

all depend upon the minimal rationality conditions, with a set of the words within these schematic conditions (some, not too many, etc.) interpreted in a cluster of different ways. Perhaps this provides as much unity as we need?

Given the interpretationist's purposes, however, this reply will be unsatisfactory if the lack of a single understanding of these terms creates problems for the process of interpretation. The question here is whether we need to settle on a particular interpretation of 'some' etc. before interpretation can proceed, and if so how we are supposed to do this.

The standard to which we hold and which we expect from a given thinker clearly depends significantly on our previous interactions with the creature, and our knowledge of its capacities and limitations. For the interpretationist, then, how to understand 'some' etc. in a given case cannot be decided before any interpretation has taken place. The interpretationist must hope that we can settle on some understanding of 'some' *during* the process of interpretation, with this understanding determined holistically along with everything else. In other words, the interpretationist must recommend that we find out which understanding of 'some' etc. allows us to produce interpretations, and then use that understanding. If this hope is warranted, then we can give a univocal account of rationality and psychological concepts in terms of what is needed for interpretation. 'Some', for example, will mean 'at least enough for the purposes of interpretation.'

We might worry that sometimes, for one and the same creature, we will be able to produce different interpretations, using different understandings of the problematic terms. We would then once again run into indeterminacy. A possible solution to this would be to say that 'some' etc. must be interpreted to include as many as possible, such that the creature can still be counted as minimally rational overall. There could then be a general presumption in favour of counting a creature

as getting things right whenever possible, and some justification would be required for counting a creature as wrong on a particular occasion.[94]

This would then also help us to address our first problem: that we have lost the assurance that we had on the consequentialist picture that a rational creature is doing those things which tend towards the satisfaction of its desires. With the presumption in favour of counting the creature as getting things right unless there's some reason to suppose otherwise, we can justifiably suppose that a subject is acting on beliefs formed in a sensible way unless there is reason to suppose otherwise. Take the Analogoid example. We see that the Analogoid was just hanging out near the sea bed, not doing very much, until a giant sea spider inched out from behind the nearby plant life and began to scuttle towards it. The Analogoid then swam quickly upwards, and the sea spider ended up standing alone on the sea bed where the Analogoid used to be. In addition, we have seen other Analogoids consumed by giant sea spiders when they did not take this action when the predator approached. Given the presumption in favour of saying that a creature is getting things right, we hypothesise that the Analogoid responded correctly to its environment and formed a belief about a predator coming towards it. This belief then allows us to explain the Analogoid's action; we can say that the Analogoid combined this belief with others formed in a sensible way and with a desire to remain alive, and in another instance of getting something right it decided that the best thing to do was to swim quickly upwards. This further confirms our hypothesis about the Analogoid's belief, and leaves us with no special reason to attribute false beliefs about other predators, mating practices, or the arrangement of wardrobes, despite the fact that these *could* also play a role in producing the behaviour that the Analogoid produced.

Finally, we may consider whether there is a lower limit to the number of things a creature has to get right, below which creatures cannot be counted as even minimally rational or as having thoughts anymore. The focus on rationality with

[94] For one set of considerations which might lead to the attribution of error, see the next section.

respect to a given thought, introduced in section 1 above, will be helpful to us here. It implies that a creature must get some things right with respect to at least many of its individual thoughts. Once again, the guiding principle will be how many things the subject needs to get right in order to justify a proposed interpretation of them.

I conclude, then, that the notion of minimal rationality, if it can be suitably developed to include conditions on the upstream connections of thoughts, provides a promising candidate for use in the interpretationist's Rationality Claim. But does it give us everything we need in a rationality claim? Does it really enable us to give good interpretations, or predictions? Cherniak himself thinks not, and adds something to the conditions above in order to enable the predictions he wants: a feasibility ordering, which is, roughly, an account of the comparative difficulty a creature has in making the various possible transitions in thought. I show how something like a feasibility ordering can be added to our account in the next section.

3. Patterns

Minimally rational agents must perform some apparently appropriate actions and inferences. However, on Cherniak's account, there are no particular actions or inferences that a creature *must* perform, (or, presumably, inconsistencies they must eliminate) in order to be minimally rational. For example, he says 'it cannot be the case that a minimal agent is able to make no inferences, but the agent can be unable to make any particular one.' (28) This creates a problem for the person who wants to interpret or predict a subject: in any given situation, should you take the subject to be doing the thing they are succeeding at, or unsuccessfully trying to do something else? Perhaps, at least in the case of interpretation, this problem is surmountable through employing the strategies suggested in the previous section for settling on an interpretation of 'some' and similar words. It is not obvious that sufficiently determinate interpretation will be always be possible, however; and the strategy

proposed does not provide any help in predicting the mistakes of a subject, which is something we sometimes seem able to do.

This section considers what Cherniak proposes we add to his minimal rationality conditions in order to solve this problem, and then argues against three objections to the idea that an interpretationist can employ his solution.

In addition to the minimal rationality conditions above, Cherniak thinks that we need information about which inferences a subject is likely to succeed and fail at. It is obvious how such information would be useful in interpretation, and also clear that it would improve the situation of the person who wants to predict a subject.

Cherniak therefore suggests that we combine the assumption of minimal rationality with information about the particular reasoner we want to predict or interpret. Most importantly, he suggests that we need a weighting of deductive tasks with respect to feasibility for a given reasoner, and that we need information about the memory capacities of that creature. The question is then what such information looks like, and how we are to obtain it.

Some, such as Hooker (1994) have suggested that this aspect of Cherniak's account shows that our conception of rationality for a creature cannot precede an empirical study of that creature, so that 'philosophical theory must ultimately be developed in interaction with empirical knowledge, not held aloof from it.' (204) It is not completely clear whether Cherniak does count his feasibility orderings as a part of minimal rationality, or as an essential addition to minimal rationality during prediction. Either way, however, this poses a problem for the interpretationist: it seems that either our conception of rationality must be developed through empirical work, or empirical investigation and interpretation of its results must be necessary before we can starting applying our conception of rationality; yet, as discussed in chapter 7, the Rationality Claim, and therefore the interpretationist's conception of rationality, are supposed to be needed before an empirical study of the creature's thinking can begin.

As well as insisting that empirical study of a creature is necessary in order to develop a suitable conception of rationality for them, Hooker also suggests that we may need a certain kind of information: 'It becomes essential to know the relevant architectural features of an agent's cognitive equipment in order to make reasonable judgements about attributing beliefs and values.' (189-90) This suggests that the current proposal not only requires information to be available at the wrong time in the process of interpretation, but that it also requires information about a creature that the interpretationist wants to reject as necessary for interpretation, namely information about internal cognitive architecture that is not available to everyday interpreters.

The interpretationist must therefore reject Cherniak's proposal as interpreted by Hooker. However, this does not mean that the notion of a feasibility ordering is useless to the interpretationist. I will deal first with the objection that incorporating feasibility orderings requires granting an interpreter information at the wrong stage of interpretation, and then the objection that it requires granting the interpreter the wrong sort of information.

Instead of seeing the demand for a feasibility ordering as a demand for information at the wrong time, we may instead see it as an instruction for how interpretation should proceed on the assumption of rationality: it must proceed by establishing a feasibility ordering on inferences for the subject as well as working out their beliefs and (if such there be) the meanings of their utterances. Working out what a creature thinks thus involves determining a third variable, not recognised by prominent proponents of interpretationism. This appears to be the way that Cherniak thinks of the situation, as evidenced in the following passage:

> the holistic interdependence of beliefs, desires and meanings emphasised by Quine and Davidson in fact extends to another domain. Beliefs, desires and meanings cannot be determined independently of at least a tacit theory of

> another type: a theory of the agent's cognitive psychology, of how the agent thinks. (48)

Interpretation, then, can still begin from an assumption of rationality, but this assumption involves more than suggested in the previous section: As well as assuming that if A has a particular belief-desire set, A would undertake at least some of those actions that are apparently appropriate, we must also assume that A has a feasibility ordering.

But what exactly does a feasibility ordering amount to? If it involves information about the internal workings of a creature, it must still be unacceptable to the interpretationist. This certainly is one way to understand the notion; however, it is not the only way. I suggest that a feasibility ordering may be constructed by interpreting a subject in such a way that they display some consistency in their reasoning abilities and strategies, given certain situations and topics. Thus, a feasibility ordering will say that a creature finds a certain mental operation difficult if our interpretation of them has them performing it rarely and only when the creature has a surplus of time to make their decision, even though it would often be useful. It will say that a creature has a limited working memory if the creature performs well on problems which require it to consider a few factors, but increasingly poorly as more pieces of information need to be brought together simultaneously in order to solve the problem. The feasibility ordering for a particular creature is thus developed during the process of interpretation, using only the information that is supposed to be available to the interpreter. At later stages in interpretation, it may then be available for use in predictions.

On this interpretation of the proposal, what we add to the minimal rationality condition in order to interpret is the same for each subject: we assume that the creature displays some consistency in their reasoning abilities and strategies. We assume that there are patterns within their reasoning processes. It is clear how this might restrict potential interpretations of a creature, and guide the process of

interpretation, despite not making things as easy as being given a ready-made feasibility ordering before interpretation began. The latter, however, is not something that an interpretationist can allow. I do not know whether this revised interpretation is more what Cherniak has in mind in suggesting the importance of feasibility orderings. However, we may note that the idea that feasibility orderings etc. ought to be available to the ordinary person fits well with his emphasis on our everyday abilities to predict one another using our cognitive theory.

So, the interpretationist could add feasibility orderings into his account of rationality. I have suggested that this might make the process of interpretation easier. Now, I will consider whether this might cause a new problem by ruling out the possibility of one off mistakes in reasoning.

Suppose that Nick finds that he has swollen glands, thinks 'If I had glandular fever I would have swollen glands,' and as a result swiftly concludes that he must have glandular fever. However, Nick is in no way inclined to infer from 'All cats are mammals' and the fact that he is a mammal to the conclusion that he is a cat, no matter how little time he is given to consider the argument. Indeed, he knows that that would be invalid because it would involve affirming the consequent, and he has recognised and criticised instances of affirming the consequent on many different occasions. What then should we say about Nick's patterns of reasoning? That he has a tendency to affirm the consequent regarding matters of his own health, but a tendency to reject that operation in other matters? What if, despite thinking about other health matters, Nick never affirms the consequent in such deliberations? What if this really is a one off mistake? If it is, then it won't fit into any pattern, and one might worry that that is a problem for the current account. In chapter 8, I presented the worry that the Rationality Claim could not cope with the systematic mistakes in reasoning that humans seem to make. Here, we are having the opposite problem.

The simple answer to this is that a subject's feasibility ordering, and their general reasoning and thinking capacities, must be manifested in discernable, but

not perfect patterns in the reasoning of a creature. Thus, we should make use of the idea of imperfectly instantiated patterns most famously used by Dennett. In his 'Real Patterns' (collected in his (1998)), Dennett explains this idea using an analogy to patterns that we might see in an array of dots, despite 'noise' being present in the pattern. However, this idea has been criticised, for example by Heal (2003). I will therefore show that the current use of the idea does not suffer from the problem Heal identifies.

Heal's objection is that, given the demandingness of Dennett's notion of rationality (she takes him to endorse something like the Standard Picture), our behaviour could not be said to instantiate the relevant pattern at all. This is a way of expressing the central criticism of the Standard Picture that I endorsed in chapter 8: we humans fall so far short of perfect rationality that the notion is of highly dubious use in understanding us. However, here we are not asking for the instantiation of perfect rationality. We are not asking for the instantiation of any particular pattern at all: just *some* pattern for each individual, to be determined by what each individual does. And any creature with any sort of moderately consistent reasoning practice will realise some pattern in their behaviour to a much greater degree of accuracy than any finite creature realises perfect rationality. I therefore think that the interpretationist could insist on the existence of patterns within thought while also allowing for some mistakes, or indeed successes, which did not fit into the pattern.

In sections 5 and 6, I consider whether we need to add anything further to the account of minimal rationality given so far in this chapter. However, given just the features described so far (i.e. sometimes getting things right and exhibiting patterns in reasoning), we can show how adopting this notion of rationality allows us to respond to facts about human limitations and the results of the heuristics and biases program, as I show in section 4.

4. Back to heuristics, biases and limitations

Since on this account being rational only involves doing *some* of those things (performing actions, making inferences, removing inconsistencies) that are apparently appropriate, some mistakes cannot disprove the claim that a creature is minimally rational. Moreover, systematic mistakes of the creature may form part of the pattern that a creature employs, and can therefore contribute to one of the conditions of rationality. Likewise, since being rational does not require removing all inconsistencies and deducing all consequences of beliefs, the fact that we cannot do this does not show that we are not rational.

The interpretationist who endorses MRC can join the consequentialist in saying that sometimes using a fast and dirty heuristic is actually a sign or an exercise of rationality: namely, in cases where a creature has decided to use such heuristics, because the use of heuristics is apparently appropriate given their beliefs and desires.

However, the interpretationist may also explain why we consider such patterns to be counter-normative, and our limitations to be unfortunate. On this view, the rules of the Standard Picture represent our engagement in a practice of discussing and evaluating reasoning. They are therefore the rules we think that it would be best for us to follow, if we had the capability, and if we would not lose other more valuable things by doing so. In addition, as some of our best ideas about what good reasoning should involve, we must also often use some of these rules to decide whether or not a creature does, sometimes, get things right.

5. Additional achievements?

Part III has been concerned with explaining the sense in which a thinker must be rational, and the initial justification for saying that a thinker must be rational was that this was surely necessary if thinkers were to be interpretable through an attempt to understand them in terms of their reasons. However, so far in Part III, I have

focused on how thoughts needed to be related to each other, and to action and environment, rather than on the idea of thinkers as possessors of and actors upon reasons. Now that we have a suggested account of rationality in terms of the former considerations, we must ask how the latter fit into the picture.

One option is to say that getting some things right and exhibiting patterns in transitions between candidate thoughts (etc.) is enough to make a creature a thinker, with its own point of view, its own reasons for doing things, and the ability to respond to and act on those reasons. We should note that the fact that exhibiting patterns is a part of this conception of rationality makes this claim a little more palatable than it might otherwise have been.

If someone responds to a reason, they must have the capacity to do so, and having this capacity can be exhibited by the person acting in accordance with that type of reason on more than one occasion. On the other hand, if a person is as likely to act against a type of reason as in accordance with it on any particular occasion, then this casts doubt on the idea that they respond to these reasons even on those occasions where they do act in accordance with them. For example, imagine a person moving chess pieces around a board. If they always move their knight two squares along one axis and one square along the other, then we will be inclined to say that they are responding to the rule for how knights move in chess. If they sometimes move their knight like this, but on other occasions move it in all sorts of different ways, we will be less inclined to say they are acting *on* the rule, even on those occasions when they do happen to get it right according to the rules of chess.

The fact that a person moves their knight in the right way on many occasions does not prove that they are responding to a rule, as is shown by an example from Child (1994). Child imagines a case where a person appears to play chess extremely well, and we are initially inclined to attribute beliefs such as 'my Queen is in danger' and 'it is worth sacrificing this pawn' to her, which we then take her to use as reasons for her moves. However, we then find out that she is looking up what

moves to make using some form of giant look-up table, and that when questioned she has no idea why the moves she makes are a good idea, nor even that what she is doing is playing a game. This is a case in which we could attribute certain thoughts, and their attribution would be useful in certain ways, but such attributions would be, on reflection, unsuitable. There is a pattern to what this woman is doing, and she often gets things right, but that is all there is: attributing thoughts about chess and how best to play it isn't appropriate.

However, in Child's example other attributions of psychological states to the very same woman are appropriate, and we can argue that *this* is what really makes the initial attribution inappropriate.[95] The case does not then show that a pattern where things are often right is insufficient for thought, but that the right interpretation of a subject can sometimes depend upon accounting for a broader pattern than we initially consider. The idea that being minimally rational in the sense developed so far is enough to make a creature a thinker with its own reasons has not, then, been shown to be implausible.

Nevertheless, some will find the claim that this conception of minimal rationality suffices for thought unintuitive. They may therefore prefer to say that not just *any* pattern of the sort described so far suffices for thought, but that there is some interpretationist-friendly thing which could be added to the account to guarantee that anyone who satisfies it is a genuine thinker. There are various suggestions for what must be added to such patterns in order to produce thought. I will provide a survey of some of them, before suggesting that we do not need to choose between these various options.

One possibility is to say that patterns which are in some ways suitable for intentional description may nevertheless fail to consist of genuine intentional states because of the rigid nature of the behaviour and apparent reasoning they involve. This sort of line is pursued by Bennett (1976). Bennett suggests that the creature

[95] Cf. the case of Christian from chapter 7, section 3.

with genuine beliefs and desires must exhibit two qualities: educability and inquisitiveness. The former involves the disposition to revise the patterns in which it forms its belief-candidate states (Bennett calls these registrations) in light of its past experience, while the latter involves a disposition to seek information. This option promotes a plausible connection between rationality and the ability to learn. It could also be revised to include other ways in which a pattern might be dynamic, for example displays of creative thought and new ideas are also popularly considered to have a connection with what is required for genuine thought. Bennett gives behavioural conditions for being educable and inquisitive, and so these would be easy to fit into the interpretationist account.

Another suggestion may be taken from Heal (2003). Heal suggests that being rational involves a 'grasp on the high level and general notion of there being better and worse in inferential transitions between thoughts' (237) which is displayed in 'effective engagement in a… practice… of asking about and assessing inferences or reasoning.' (Ibid.) In addition to this, Heal suggests that particular thinkers have an inferential outlook, shown in the actual transitions the thinker makes, as well as the examples of good and bad reasoning she would cite and any explicit theorising about reasoning.

Heal says that we should see rationality 'as a capacity, rather than a particular achievement.' (237) However, her account does require the rational being to achieve certain things: they must have a particular idea (of better and worse inferential transitions), display this idea by making a certain kind of progress, and also displaying something which looks a great deal like the feasibility orderings and patterns I suggested were important to thought in the previous chapter. Heal is not arguing that rationality could be a mere capacity in the sense I said was problematic in section 1. Rather, she rejects the ideas that a pattern could count as involving thought despite failing to develop over time, and that obeying any particular norms is necessary for rationality. We have already rejected the second idea, and we can

incorporate the idea that rational patterns need to be dynamic in some way without giving up the idea that it is patterns that we are interested in (as with Bennett).

The interpretationist could therefore appropriate Heal's ideas, and say that in order to count as responding to reasons, and therefore as a thinker, a being must be interpretable as having the idea of better and worse transitions in thoughts. They could then say, along with Heal, that

> To credit someone with the... ability to engage effectively in the practice of asking about and assessing reasons for forming beliefs and intentions... is in effect to credit that person with a capacity for coming to know at least something about this realm of norms.' (239)

Another suggestion is offered in Moran (1994). He suggests that rather than merely being subject to norms of rationality, a subject must *intend* his behaviour and reasoning to conform to such norms. It is *this* which he thinks allows us to say that

> Invoking rationality in the first place thus commits the interpreter to a notion of justification, which means justification *from the standpoint of the agent and his thought*, rather than from the standpoint of the explanation or the explainers.' (1994: 168)

From an interpretationist perspective, Moran's suggestion then amounts to the idea that a candidate thinker must display a pattern of behaviour which allows us to attribute to them a particular kind of intention.

Finally, Davidson also suggests that a particular mental state possessed by a creature is key to their being a thinker: he suggests that a thinker must grasp the distinction between how things seem and how they really are. He describes this idea variously, saying for example that 'Someone cannot have a belief unless he understands the possibility of being mistaken' (1984: 170) and must 'command… the

contrast between what is believed and what is the case.' (2001:105) Davidson links this to creatures having the idea of belief at their disposal, and claims that this idea can only be acquired and used by creatures that are in linguistic communication with others. However some, such as Verheggen (1997), have suggested other circumstances in which one might gain and apply this idea, namely cases when past experience conflicts with present experience, or cases when the information from one sense modality conflicts with another.

The idea that Davidson suggests as necessary for thought seems to be connected to the ideas and thoughts which Heal and Moran pick out as important and the behaviours which Bennett sees as important: it is in light of the idea that there is an objective truth which one might be right or wrong about that the idea of better and worse ways of finding out about that truth, and the intention to use certain ways of finding out about it make sense.

All of these proposals have some promise, and, as presented thus far, all also face some problems. For example, one might worry that Bennett's proposal unreasonably rules out the possibility of a thinker who already knows all that it needs to know or has the capacity to remember about its environment, and which is able to achieve all of its purposes using this information. With respect to the second proposal, one might agree that the creature to whom we could attribute the thoughts and capacities mentioned by Heal would be an eminently suitable candidate for the title of 'thinker'; however, one might worry that her requirements are too demanding to be necessary. In particular, if thinkers must engage in a practice of asking about and assessing reasoning, then only language users can be thinkers. Although congenial to a few (such as Davidson), this constitutes a highly controversial claim which does not accord with ordinary usage of psychological terms.

Various thought experiments might be used to test these proposals, but different people have different intuitions about when and how widely we want to

apply the concept of thought, and will be attracted to different proposals as a result. I will suggest that a definitive choice between the options canvassed is in fact unnecessary.

We might say that static patterns of behaviour which allow us to give a reasonably determinate interpretation of a creature (because it is possible to interpret the creature as getting a reasonable number of the necessary sorts of things right) are linked to a very basic notion of a point of view. They are linked to the idea of a thing which can prosper or fail to prosper, and we can see such a being as having projects to promote and thoughts which guide its behaviour in the fulfilment of those projects.

Learning and creative thought are clearly related to the notions of rationality and responsiveness. However, why should we insist that they ground even the most basic notions of responsiveness and perspective, rather than being important extensions of them? We could instead conceive of them as adding something to the basic notion of a point of view by being part of a less basic (and perhaps also more important) notion. Self-awareness and explicit consideration of norms constitute another major development to the notion of a point of view. However, they seem too demanding to cover all that we seem to mean when we talk about thought.

The task of showing that one of these proposals provides what we need in order to be justified in applying the concept of responsiveness to reasons holds little promise, in my opinion. I suggest that this is instead a case where it is not appropriate to make a decision and say that one notion of rationality, and therefore thought, is the correct one to use, even given the caveat that we are talking to interpretationists. Rather, there could be different potential notions of rationality and thought which are better for different purposes, and thus far it is unclear why any decision must be made between them.

This then brings us to the question of what the purposes of interpretation are, an issue which I address briefly in section 6.

6. The purposes of interpretation

Interpretation involves attributing reasons to a subject in an attempt to gain an understanding of how things look from their point of view. But what is the purpose of this exercise? According to Dennett, the intentional stance is just one of three stances that we might adopt in order to fulfil the same aims of prediction, explanation and control. One of the other stances is the physical stance, the stance taken up by the scientist. Dennett therefore seems to suggest that interpersonal interpretation has the same purpose as scientific theorising.

This chimes with a traditional view of the purpose of folk psychology, taken for granted by some of its supporters and critics alike.[96] However, it has been challenged in recent literature.[97] I do not have space for an in depth consideration of the potential purposes of interpretation, but will briefly consider what additional/alternative purposes it might have and how this might affect interpretationism by considering and extending Heal (2003)'s suggestions. I will suggest that this provides a way to show the connection between the different potential notions of rationality and thought suggested in the previous section.

Having rejected the idea that our interactions with people are of the same basic kind as our interactions with inanimate objects, Heal (2003) suggests that interpretation may underpin kinds of interaction between people which were overlooked or distorted on the account which saw interpretation theory as having the same role as science. She suggests that such interactions are many and various, but picks out two as particularly important: inviting and participating in joint consideration of a topic, and dealing with disagreement and excuse. The first is supposed to make sense precisely in light of the idea that oneself and others are responsive to norms, while

[96] In addition to Dennett (1987), see Churchland (1981) and (1991), Gopnik and Meltzoff (1997), and Jackson and Pettit (1993) for examples.
[97] See Heal (2003), Morton (2003) and Knobe (2003) and (2006) for some examples.

the second involves an attempt to find an account where although a person has made an error, arriving at that error nevertheless makes sense, and was an exercise of rationality. In light of such examples, Heal claims

> Other people are not devices which we try to operate, endeavouring to cause them to do this or that useful manoeuvre. Rather they are fellow human beings with whom we talk, with whom we cooperate on shared projects, from whom we ask help when we are muddled and with whom we seek to forge a jointly created and growing understanding. (245)

The enterprises above could perhaps be redescribed in terms of the language of prediction and control. However, the idea is that this would be a misleading, less helpful way to describe them. There is an analogy here with the way in which any action can be redescribed to accord with a thesis of universal egoism. Even if such redescriptions are possible, they introduce unnecessary complications or threaten to make the thesis vacuous.

The interactions Heal mentions seem important, and she explicitly says they don't exhaust the field. However, it is noticeable that both of her examples seem particularly at home among language users. We might therefore try to mitigate the impression that Heal's proposal applies only to the interpretation of language users by giving a more general statement of the proposal. For example, we might say that the purpose of interpretation is that the interpreter take up the point of view of the subject, with the result that additional interactions, projects and relationships become possible.

The taking up of another's point of view may be realised to very different degrees. In particular, our ability to take up another's point of view is affected by the difference between our knowledge and capacities and those of the subject. A particular problem therefore arises with non-linguistic creatures. The sorts of words and sentences we are inclined to use to state the content of psychological states are

those which we use to describe our own mental states, and the world as we see it from our point of view. They gain their meaning by the place they have in our languages and lives, and as a result imply more than we want them to when we apply them to non-linguistic creatures. However, we may see the degree to which we can take up another's point of view as partly a function of the strength of the notion of a point of view that is appropriate to the creature in question. Even if the psychological ascriptions we can give in our language are not completely appropriate in all cases, they may still capture something important, and enable the fulfilment of purposes that are important to us. The fact that they do so is shown by the way applying them satisfies the second component of the condition above: certain interactions, projects, etc. do become possible only when we conceive of certain non-linguistic animals as a variety of thinker. It is perhaps the suggestion that new relationships become possible through interpretation which differentiates its purposes most clearly from those which can be achieved by doing science alone.

The interpretation of non-linguistic creatures is discussed further in the next chapter, where I argue that although we can attribute thoughts to them, the sorts of thoughts we can attribute may be limited in interesting ways. The interpretation of beings who are different from us in other ways, including those who are much more intelligent than us, is also an important issue. It is broached briefly in chapter 12.

There is another way in which the consideration of the purposes of interpretation may be useful to the interpretationist: it might be used to further develop his account of when the ascription of thought is appropriate. Given the suggestions of this section, we might argue that Dennett is correct that it is what you could gain from using the intentional stance that makes it appropriate to use in a particular case.[98] However, we might say that what you gain is not what Dennett suggests. When the intentional stance is truly appropriate, you do not gain an ability to explain and

[98] See his (1989) especially p. 22-27.

predict which you could also get by using the physical stance if you were smarter. You gain something, and make possible projects, that are not available from any other perspective. Further discussion of the purposes of interpretation, then, may place extra constraints on what counts as a thinker.

7. Conclusion

Something like Cherniak's account of minimal rationality appears to combine elements from both the Standard Picture and the Consequentialist picture to provide a conception which makes rationality obtainable for ordinary humans, and which could be useful in interpretation. This notion requires the rational creature to exhibit some achievements (they must get some things right, and they must display some pattern in reasoning) and it provides the basis for an interpretationist account of rationality.

Whether more should be added to this account of rationality is to some extent a matter of taste. I propose that there are a group of notions of rationality in the area, with the one which adds nothing further connecting to a very basic notion of thought and the idea of a point of view, while others connect to more substantial versions of these notions. These notions of thought and rationality are connected by their relationship to a central purpose of interpretation: taking up the point of view of a subject, with the result that additional interactions, projects and relationships become possible. More elaborate interactions, projects and relationships of course become possible as we ascend to more complex notions of a point of view and of thought.

There are various ways in which the account presented in this chapter requires further development. However, I hope that it provides the framework for a more detailed interpretationist picture of rationality. The theory presented so far has also enabled us to answer the challenge that human limitations and the results of the heuristics and biases program pose for the Rationality Claim, and in the next chapter

I argue that it allows us to say something about the interpretation of thoughts that seem to have little to do with rationality.

Chapter 11 – Non-Rational Thought

In this thesis, I have presented interpretation as a scheme for gaining a particular sort of understanding of others. This form of understanding has involved making sense of another, attributing reasons and responses to those reasons, seeing how things look from another point of view and working out why actions made sense from that point of view. I have claimed that through this process we can attribute intentional states to others, and that it is within this scheme that thoughts belong.

In Part III, I have also argued that this process allows us to attribute instances of what, on some understandings of the term, we might call irrationalities. We can attribute thoughts to creatures even though those thoughts involve one-off mistakes in reasoning, systematic mistakes in reasoning, and even purposeful employment of logically invalid procedures for coming to conclusions. The idea was that as long as a creature displayed patterns in its reasoning, and got enough of the right sort of things right with respect to each thought, we would be able to identify the thoughts. For example, take Nick's mistaken conclusion that he has glandular fever in chapter 10. He made a mistake in inferring it from the beliefs that his glands are swollen and that glandular fever includes swollen glands. But suppose he then decided on the basis of the unfounded belief to look up treatments for glandular fever on the internet, as a result of a very sensible chain of thoughts. This may then be one of the things that lead us to attribute to him the belief that he has glandular fever, even when we cannot see it as the outcome of a valid argument on his part.

But perhaps the focus on reasons will engender a feeling of unease. All the thoughts we have looked at in this part of the thesis have been thoughts involved in traditionally rational kinds of thought processes: in trains of reasoning, both theoretical and practical. However, these are not the only kinds of thoughts we have: not all of our mental life consists of a kind of argumentation. What then should we say about other kinds of thoughts: is there a problem in determining their content,

because the Rationality Claim which is supposed to help us determine content has nothing to do with them? Has the focus on reasons and rationality created a theory which leaves out a great swathe of our thinking?

The primary aim of this chapter is to consider two kinds of thought (dream thoughts and imaginings) and one kind of thinking (associative thinking) which do not seem to have very much to do with reason, and to suggest how these can be fitted into the suggested scheme of interpretation.

Section 1 introduces the issue above as a second problem with dreams. I suggest that dreams may involve imaginings, but in section 2 I suggest that this merely widens the scope of the problem. I consider and reject an account which tries to connect imaginings and rationality by seeing all imaginings as a kind of intentional action, done for reasons. It is more plausible to suppose that some imaginings are produced through a different process: association, discussed in section 3.

In section 4, I present an account which allows us to identify and attribute these problematic thoughts to any language-user. Section 5 then discusses whether such thoughts can also be attributed to creatures without language, and suggests that some can, although they will be less complicated and may also exhibit greater indeterminacy. Finally, section 6 admits that my account makes interpretation explanatorily incomplete. I explore the nature of this incompleteness, and use it to outline the relationship I believe the interpretationist should posit between interpretation and science.

1. Another problem with dreams

In chapter 4, the problem identified with dreams was that they seemed to be hidden from interpreters. They pose another problem as well: they are notoriously weird. Sometimes they are incoherent with what we ordinarily take ourselves to think, and

sometimes incoherent even within themselves. Alternatively, while not actually incoherent, they can seem strangely disconnected from the rest of our mental life, or from other parts of the same dream. Some have even characterised dreaming as a kind of madness.[99]

Sceptical arguments sometimes proceed from the idea that one believes things that are false while dreaming. If this idea is correct, then dreams also include beliefs which are inconsistent with our occurrent, waking beliefs and with standing beliefs which we might suppose to persist during sleep. Such dream beliefs are also usually false and disengaged from our usual belief forming strategies.

Perhaps the problem would be easier to deal with if we could at least reject the idea that dreams involve beliefs. We could suggest instead that they are only cases of imagination. This is not a new idea: it is suggested in Walton (1990), and explored in some detail by Sosa (2005) for the purpose of defeating scepticism. However, if this suggestion is to help the interpretationist, we must be able to give an interpretationist account of imagining. Otherwise, we have merely identified another problem for interpretationism. And indeed, imagining does seem to be another sort of mental state which has little to do with rationality.

2. Imagination

Very little has been said about imagining by interpretationists. This seems a significant oversight, since imagining sometimes involves intentional states, and seems to be importantly different from the sorts of states interpretationists have focused on, namely those which play the biggest role in reasoning. For example, we may compare imagining with belief: belief is supposed to be somehow constrained by what is true about the world; its formation is supposed to be regulated by principles; and it usually has effects on action. Imagining, in contrast, seems

[99] See Hobson (2005), chapter 7.

remarkably free, and although we may act out what we imagine, we can imagine without doing so. Indeed, according to Hume, 'Nothing is more free than the imagination of man'. (2007: 47) It may therefore seem as if imaginings have nothing to do with rationality.

However, Hume's conception of human freedom does not involve the possibility of doing absolutely anything, regardless of whether or not you have any desire or reason to do it. Rather, Hume thinks that a person is free when they are able to control what they do in accordance with what they want. This suggests a way of explaining the freedom of the imagination without saying that imaginings fall under no rational constraints whatsoever: imagining can be free and done for a purpose. In the rest of this section, I consider this proposal, but argue that it does not give a plausible account of all imaginings.

One obvious purpose of imagining is wish fulfilment, but there are many other candidates as well. Kendall Walton suggests the following purposes for waking make-believe: '[it] provides practice in roles one might someday assume in real life… it helps one to understand and sympathise with others… it enables one to come to grips with one's own feelings… it broadens one's perspectives.' (1990: 12) We could add more: that we might imagine to crystallise for ourselves what we want to achieve, or equally what we want to avoid, or to fix things more securely in our memory, to escape from our current situation, or simply for pleasure. If they are explained in this way, neither imaginings in general, nor the thoughts which occur within dreams, need to be seen as irrational or disconnected from reasoning.

A striking feature of this account is that it proposes to characterise imagining as an intentional action, apparently done in accordance with beliefs and desires. This seems plausible for some cases of waking, purposeful imagining. For example, it seems reasonable to explain a day dream in the following way: Nathan is bored, and desires to alleviate this boredom. He believes that imagining being a superhero and swooping in to catch Amanda as she falls from a tall building will be pleasurable. He therefore performs this imaginative exercise.

However, other cases of imagining are more resistant to this sort of treatment. Dreams provide one difficult case: if dreams include imaginings, then on this account they need to be at least partly composed through intentional actions. But it seems far more problematic to posit beliefs and desires about how to achieve ends through dreaming, and so to say that genuine agency is involved in the production of dreams. Many cases of waking imaginings also resist this sort of treatment: instances of imagining often seem to be quite spontaneous, their content unexpected, and their effects predictably unhelpful. Consider for example the person who stands on the edge of a cliff, and who engages in a sudden and detailed imagining of deciding to step over the edge, which makes them feel frightened and dizzy.

Although I borrowed from Walton in suggesting the purposes that a person might have in imagining certain things, Walton himself does not posit the purposes he mentions as purposes *of* the imaginer. When he says that children's make-believe games help them to prepare for assuming roles in later life, he is not saying that children have any *beliefs* about what will prepare them for their future. The suggestion is certainly not that children explicitly plan such games. Indeed, it seems very plausible that it is unnecessary for children to think explicitly about their future in order to play, and that they might not be able to say why they were playing particular make-believe games even if we asked them. I think that we should agree with Walton on this, and that we should also say that adults who use their imaginations *need* not have particular purposes in mind when they do so. This means that we should not see all imaginings as intentional actions, and so should not connect them to rationality in that way.

Perhaps, however, we can still see imaginings as the result of reasoning if, rather than saying that creatures have conscious purposes in imagining, we attribute unconscious purposes to them. So, for example, we might say that the man who imagines deciding to step over the edge of the cliff must have an unconscious death wish, and we might posit other unconscious states to explain the content of dreams. We would then also need to say that such states were difficult to access and

influence (at least in many cases, including the case of those thoughts which cause dreams). This takes us into murky waters involving issues in the philosophy of psychoanalysis. Positing such unconscious thoughts and chains of reasoning is quite a drastic step, and I will argue that it turns out to be unnecessary in order to solve the problems of this chapter.[100]

Instead of positing unconscious desires to explain dreams and some imaginings, many would feel happier saying that dream thoughts and imaginings can be produced through a process of association. This process is the subject of the next section.

3. Association

In cases of association, one thought appears to cause another, but the first is not a reason for the second. An example would be if I were to think 'Herbert Hoover was very unfortunate to become president when he did', and because of this my next thought was, 'I need to vacuum the sitting room.' It seems clear how a process of association might produce many of the features of dreams, and Hobson suggests that association plays an important role in determining the content of dreams in his *Dreaming: A Very Short Introduction* (2005).[101]

As the example shows, it is not only dream thoughts and imaginings which may be produced through association. We may therefore worry that association widens the scope of our problem even further, to cover even more thoughts. However, this worry is unjustified. Although the belief about needing to vacuum the sitting room is caused to occur at the time that it does by another belief which isn't a reason for it, it is nevertheless also caused by other thoughts that are reasons for it (a memory of what my sitting room carpet looks like, and a desire for it to look cleaner). It may also go on to have rational downstream connections to an intention

[100] There might be other good reasons for positing such states, but I will not discuss them here.
[101] See, for example, p22.

to vacuum the sitting room when I get home, and the action of doing so. We don't get a 'rational' explanation of when the thought occurred, but otherwise it does fit nicely into the project of making sense of another into terms of reasons.

Thus the problem, of thoughts with intentional content which nevertheless seem disconnected from reasoning, does not seem to spread to all thoughts produced through association. But imaginings and many of the thoughts within dreams remain a problem. If they are indeed produced through association, and they need not possess other either upstream or downstream 'rational' connections, then it is hard to see how they can fit into the scheme of interpretation that I have proposed. To solve the problem, we must return to an argument that I gave in chapter 10, section 1.

In chapter 10, I argued that the interpretationist should conceive of rationality as involving achievements. However, I argued for this by looking specifically at the case of belief, citing a conceptual difficulty in conceiving of a belief which had neither the right upstream nor downstream connections to other thoughts. I suggested that beliefs just aren't the sort of thing which you can have, but not have formed appropriately nor use to guide your thought and action even when you can and should do so. We must now question whether these sorts of considerations apply to all thoughts.

4. Verbal reports

Imaginings, entertainings and so forth do not seem to have the same strong connection to the norms of rationality as do beliefs. This has been presented as a problem for the interpretationist who wishes to fit them into his theory. However, it also suggests that such thoughts may not be vulnerable to the argument from chapter 10. For such thoughts, then, we have no argument for the necessity of actual rational connections to other thoughts. Even in the case of beliefs, the argument relies on the premise that there will be occasions when the belief should make a

difference to one's other thoughts and actions. This seems extremely plausible for many beliefs, but perhaps it is not plausible for all of them. If beliefs do occur within dreams, then these seem good candidates for beliefs which one never has occasion to connect with other thoughts in a rational way.

Thus, in the case of the thoughts in question, the focus on actual achievement in the domain of rationality seems less important. A mere capacity or disposition to employ such thoughts as part of a rational pattern of thought and action may be sufficient here. Still, we need to show how such thoughts can be fitted in to the project of making sense of another creature, and how this can allow the interpretationist to account for their content.

The suggestion of this section is that we return to some of the ideas presented in chapter 4. There, I argued that the possibility of a dream resulting in a dream report was enough to make dreams interpretable. This possibility can also help us to understand how the attribution of dreams may fit into the project of interpretation.

If a person reports a dream, imagining or other thought produced through association, the interpretation of the thought in question will then have a place in the overall project of interpreting the whole person, just because the report occurs in a language which can only be interpreted due to many interactions with and observations of the person in question. A variety of actual connections between thoughts then remain important: namely the relationships between thoughts which can be expressed using the same words and which involve the same (at least basic) concepts. For example, on this account you could not have a person who only had thoughts about their village and the people in it while awake, but then had a dream which involved thoughts about quarks. This must be impossible because, due to the fact that there is no connection between what his dream is about and what his waking thoughts are about, the interpreter who knew his language from interpreting his waking behaviour would have no chance of interpreting a report of his dream.[102]

[102] The proposal therefore endorses the importance of the relationships between thoughts picked out as important by type C holists (see chapter 4).

Reasons also remain important on this view: the thoughts in question are counted as having a certain content due to their potential to result in reports, which may be seen as caused by a belief (formed in whatever way we find out about our own thoughts) and a desire to communicate this. Reasons also still play an essential role in the interpretation of the whole person and their language.

This account suggests that dreams are importantly dependent upon the existence of certain waking thoughts and actions. Dreams can count as thoughts only because of the role they play in a system which includes uncontroversially accessible and rational mental states. Both dream thoughts and the other thoughts discussed in this chapter depend upon thoughts which stand in canonically rational relationships to one another.[103]

Given this account, the thoughts being considered in this chapter can fit into the project of making sense of another creature, and we can give an interpretationist account of why they count as the mental states that they do, without needing to posit unconscious states which can provide reasons for them. This does not mean there is no reason to posit such states; it merely removes one reason for doing so.

5. Non-linguistic creatures

I have argued that although dreams and imaginings may seem to have nothing to do with rationality, the connection the interpretationist needs between these thoughts and rationality can be found due to the potential for thinkers to report such thoughts. But what about creatures which cannot give such reports, because they do not have a language? Must the interpretationist say that such creatures (including some humans) cannot have thoughts during dreams, nor imagine certain things? This

[103] Although this need not be a one-way dependence: one could hold a position according to which there was some sort of mutual dependence, perhaps because thoughts produced by association provide necessary material for the canonically rational processes to work upon, or form part of the canonically rational thoughts.

may seem particularly implausible in light of the fact that all mammals show brain activation during sleep, and young mammals, including pre-linguistic humans, exhibit the bodily signs which, in human adults, we might take to be signs of dreaming, such as movements and facial expressions.[104]

I will start with the issue of dreams. Initially, there seem to be two potential interpretationist-friendly replies to this problem: that languageless creatures do not have dream thoughts, but might have other, non-intentional mental states while sleeping; or that languageless creatures can have dream thoughts, but that these are massively indeterminate. I outline each, but do not decide between them. I then suggest that the situation is a little better in the case of waking imaginings, but that we must admit that the imaginings of languageless creatures are more restricted and less determinate than the imaginings of those with language.

The first reply to the issue about animal dreams calls on the fact that, as outlined in Part I, the interpretationism I am interested in is an account of thought content, not an account of all aspects of the mental. The kind of interpretationism that I am interested in therefore does not need to say that, in languageless creatures, brain activity, movements and facial expressions during sleep do not correspond to *any* kind of mental activity. They may allow that such creatures can have a kind of experience during sleep, and can have emotions, depending on the accounts they want to give of such states. However, they might say that the kinds of states that they are most interested in (those I have been calling thoughts, and which have intentional content) are ruled out because they could not possess the rational connections to other thoughts, the environment and action the possibility of which is required for belief. They would then say that non-linguistic or pre-linguistic creatures may experience states related to our dreams, but that the interpretationist

[104] Hobson (2005: 65-6).

shouldn't attribute to them all the features that our dreams can have, including the features with which he is most concerned.

Alternatively, the interpretationist could say that languageless creatures can have thoughts during sleep, and that these have the content they do because of their connections to the creature's previously experienced environment and movements during sleep. However, they will at least sometimes have to say that these thoughts are *highly* indeterminate. How indeterminate a creature's dream thoughts will be depends on how different the creature's patterns of behaviour are in different situations (the more differentiated the creature's waking responses to different situations, the more thoughts we can differentiate between on the basis of movements during sleep). Still, for any creature without a language, it seems likely that their dream thoughts cannot be as determinate as those of a creature which does have a language and can report on their dreams, nor even as determinate as many of the thoughts we can attribute to it during its waking life. Perhaps they will be so indeterminate we would prefer to refrain from counting them as thoughts at all, in which case this second option collapses into the first. Or perhaps this will only sometimes be the case, in which case we should give different answers for different creatures, depending on the richness of their behavioural repertoire.[105]

Next, consider wakeful imaginings. Potential rational connections may certainly exist between wakeful imaginings and actions other than self-reporting, for example in the case where one acts out the things one is imagining during play. In such cases, there is more for the interpreter to go on than in the case of dreams,

[105] The choice between these two options, if such there be, will also be connected to an issue raised, but not decided, in chapter 3, section 4: the issue of the weight we should give to sleep behaviour as additional criteria (on top of dream reports) for the application of our concept of dreaming. If we say that sleep behaviour has as much weight as dream reports in a decision on whether to apply the concept of dreaming, and that this is sufficient for the application of the concept, then I think we should go with the second option. On the other hand, if we decide that sleep behaviour could count as evidence, but never conclusive evidence of dream thought, then the first option would be more appropriate. As in chapter 3, I won't decide between these two options.

because the creature will be able to exhibit a greater range of behaviour, and will be able to interact more with their environment as part of the exercise. In such cases, the second suggested response to dreams seems appropriate: we may say that we can attribute thoughts, but that they will probably always be less determinate (and also less complicated) than the wakeful imaginings that language users can have, and may also be less determinate than some of that same creature's other thoughts.

As I suggested in chapter 10, I once again conclude that the words we use to describe the mental states of languageless creatures are in some ways inappropriate, because their primary place is within our practices of describing our own thoughts and those of other humans. Still, they capture something important; a similarity between ourselves and languageless creatures, and thereby underpin the possibility of new interactions, such as playing together.

I do not think that we need to say that all imaginative states are ruled out for creatures without a language. However, such states are certainly limited by the sorts of things that can be shown through the actions of the creature. For example, we almost certainly have to deny that an orangutan could make up certain sorts of story, which included the thoughts of protagonists etc. However, this doesn't seem unreasonable. Putting restrictions on the thoughts said to be possible without language is not an unusual commitment even well outside interpretationist circles.

6. Explanatory incompleteness

I have shown how states such as dream thoughts and imaginings can be linked to rationality as I have characterised it. The connections picked out as important are sometimes only potential, and they are downstream of the thought in question. This latter feature in particular has an important consequence: it means that I have not shown that we can *explain* such states in the way appropriate to our scheme of interpretation. On the account I have given, there need be no reason attributable to

the person, nor feature of thought in general, which tells us why a person had the particular dream, or made up the particular story, that he or she did. There is therefore a kind of explanatory incompleteness to the view of thought and interpretation that I am offering.

This brings us to an issue about the explanatory ambitions of interpretation. Consider the following distinction, offered in Gardner (1993):

> On the Complete view, ordinary psychology is committed to there being in principle a full explanation in its own terms, or in terms congenial to it, for the psychological states that it cites as the proximal causes of action. On the Limited view, by contrast, ordinary psychology leaves undecided, and is indifferent to the existence of, explanation beyond a certain distance from its immediate point of application. (1993: 33)

The distinction is introduced in the context of a discussion about how to explain the traditional problems of irrationality, such as self-deception, and Gardner's main concern is arguing that an acceptable account of such irrationalities must allow us to explain, in particular cases and in terms congenial to ordinary psychology, why a person's thoughts take an irrational rather than a rational path. I have not been concerned with these forms of irrationality, and will not discuss them in this thesis. Nevertheless, one might try to import Gardner's distinction into the discussion of this chapter.

By allowing that certain states can be genuine thoughts and causes of action, even though they are not susceptible to ordinary psychological explanation, my account rejects the Complete view. It does not, however, thereby endorse the Limited view: it is not indifferent to *any* such explanations beyond a given point. Rather, it allows some unexplainable thoughts as long as they take place in the context of many explainable ones, and the question of how a state was caused and can be explained plays at least some role in the account of what sort of state it is. If

we import Gardner's distinction into our context, it therefore presents a false dichotomy.

There is a second feature of my account of interpretation which might be referred to as explanatory incompleteness: interpretation certainly does not tell us how a process such as association works. Such questions seem perfectly sensible, and they also seem to require something other than the sort of response one gets through our ordinary practices of interpretation.

Although this is a recognition of a kind of incompleteness, it is not the sort of incompleteness which worries someone like Gardner: he calls these mere 'nominal gaps in ordinary psychological explanation' (1993: 227) and says that 'they do not betoken a missing piece in the pattern of personal understanding; ordinary psychology regards memory and powers of recognition as a level of given competences.' (Ibid.)

I argue that, first, this does not mean that interpretation and interpretationism have nothing to say about such processes; second, it suggests the form of the relationship between scientific study of the mind and ordinary interpretation; and third, that we can also allow a kind of scientific explanation of the content of mental states.

First, although everyday interpersonal interpretation does not aim to given an account of how certain mental processes work, it is involved in determining what *counts* as a given kind of process. To take the example of association, our scheme of interpretation demands that such a process result in thoughts which are caused by other thoughts which do not count as reasons for them, but which possess some sort of connection (from the point of view of the creature) to the thought produced.

In addition, we should recall that interpretation proceeds (according to chapter 10) through the identification of patterns in reasoning. This includes measures of feasibility for a given creature for retrieving useful material from memory in certain situations, and also feasibility of holding certain amounts of

information together for consideration. Interpretation could also include picking out patterns in a creature's exercises of its associative powers. It is unclear whether recognition of such patterns could actually be necessary for interpretation. I will not try to resolve this issue, but maintain the claim that interpretation can allow us to describe the surface contours of a process like association in a given creature.

Interpretation, and therefore interpretationism, may therefore have something to say about mental processes like association. However, there are plenty of things about the realisation of such processes about which interpretation tells us nothing. There can clearly be scientific investigation of how such processes work and why they occur in us, and these may form a part of a scientific account of how thought is (at least usually) realised in a given species. Branches of science may also tell us how we humans came to have minds, and why this process resulted in us having the sorts of minds that we have. This might well include explanations about how we came to be prone to certain sorts of mistakes (like those discovered in the heuristics and biases program) and irrationalities (such as the traditional problems of irrationality). There may also be scientific explanations concerning the minds of particular people: different humans are more prone to certain kinds of thought than others, and we may discover explanations for why some find it more difficult to perform abstract thinking and others are prone to depression etc.

Ultimately, I think we should allow that there can be scientific explanations of why people have the particular thoughts that they do. For example, the contents of states produced through association are determined by rational connections, but they are not explained by those rational connections, as I explained earlier. It seems plausible to me that such explanations just aren't available for these states. But another sort of explanation, the sort given by science, may be available (although of course interpretationism in no way depends on or guarantees this). Even the content of a mental state may be explained in this way: sometimes evolutionary explanations seem to do just this. For example, we might suggest that thinking about eating the red sweets led Harry, through a process of association, to think about the time he ate

some red berries and was then violently sick, because connecting instances of new food consumption according to such similarities was evolutionarily useful to our ancestors. This is a genuine explanation of a particular thought with a particular content. It is just not the sort of explanation which directly determines what sort of thought the thought in question is. There will be another, compatible explanation of that which calls on a particular way of making sense of Harry as an agent with reasons and on the sort of evidence which could be available to ordinary interpreters.

I take interpretationism, then, to say that scientific investigation has the potential to tell us very interesting things about human minds in general, the minds of particular people, and particular thoughts. It just doesn't get to make pronouncements on the nature of thought in general, nor which of our mental concepts to apply in a given case. This is what gives our beliefs about thought the protection from refutation that we wanted in chapter 1, section 3.1.

7. Conclusion

The process of interpretation that I have described through Part III does allow us to attribute dream thoughts (perhaps including beliefs), imaginative states and thoughts formed through association to thinkers. Moreover, consideration of this issue has helped us to clarify the differences between the interpretation of and the thoughts attributed to language users and languageless creatures, and to say something about the explanatory ambitions of interpretation.

There are a great many questions about the process of interpretation remaining. Notably, I have not considered the interpretation of certain kinds of irrationality, including the traditional problems of irrationality and some of the thoughts which

occur during mental illness;[106] I have not established whether there are some kinds of irrationality which the interpretationist cannot countenance; and I have not established whether the interpretationist needs to posit unconscious mental states.

However, the account of rationality and interpretation developed in this thesis provides a framework for the consideration of these interesting issues, and I think that it should form the basis of any interpretationist account.

[106] The idea that these may cause problems for interpretationism can be found in Gerrans (2004) and Bortolotti (2005).

PART IV: CONCLUDING REMARKS

In chapter 1 considered the following question:

> Are there particular sorts of internal organisation which a being must have in order to count as a thinker?

I then offered three reasons why one might find the answer 'No' attractive. I have tried to develop interpretationism with these reasons in mind.

In chapter 12, I consider whether the theory presented in this thesis does exhibit the positive features that a position in the 'No' camp is supposed to have. There are complicated issues here, which cannot be treated fully in a concluding section. I aim to make it plausible that the interpretationism developed does exhibit several positive features in this area, and to indicate the shape of some future discussions. The conclusion will be that further development of interpretationism looks fruitful.

Chapter 12 – Fulfilling the Promises

In chapter 1, I suggested that positions in the 'No' camp would find it easier to meet certain conditions than positions in the 'Yes' camp. It would be far too large a task to compare how interpretationism fares on these issues in comparison with the large range of 'Yes' camp theories. Here, therefore, consideration is confined to whether interpretationism does appear compatible with the origins and uses of our psychological concepts (section 1); what protection it gives to our beliefs about minds (section 2); and whether it allows us to avoid chauvinism (section 3).

1. Origins and uses of psychological concepts

The first issue to consider is whether interpretationism fits with the ways we learn to use psychological concepts and how we typically go about attributing psychological states. There are certainly things to be said in interpretationism's favour here. First, it focuses on behaviour, which is definitely vital to us in learning to use psychological concepts and central in our application of them. Second, it draws attention to some plausibly distinctive uses of psychological concepts and to a distinctive sort of understanding that is involved in employing them (chapter 10, section 6). By doing this, it focuses our attention on the things which are most important to us about the creatures to which we want to apply such concepts, such as the sorts of interactions and relationships that are possible with them. However, under this heading we must also consider an issue flagged in chapter 2, section 2: that of the relationship between the method of interpretation invoked by interpretationism and our real methods of finding out about other people's thoughts.

The conclusions of this thesis so far have put us in a better position to address this question, because they have provided more details about what is involved in the

method called on by interpretationism. In particular, we have refined our understanding of the sort of evidence that is relevant to the interpretation in interpretationism, and of what it means to use the assumption that a creature is rational in this context. This puts us in a better position to work out if this is a process of interpretation that we use, or how it relates to the process we use.

I hope that the process of interpretation that I have described through the thesis sounds familiar, and seems to be a method that we do, at least sometimes, engage in. If so, this will help us to say that the process described is of relevance in understanding our method and answering philosophical questions about our beliefs about minds.

Another point in favour of there being a close and illuminating relationship between the interpretation of interpretationism and our real method is that I have avoided at least some of the idealisations of our method that would have ignored important human limitations. For example, in chapter 3 I emphasised that the important process of interpretation must not depend on information which couldn't be available to the ordinary interpreter. There is, however, a topic in this area which has received comparatively little discussion; namely how, given behaviour and environment, plus the assumption of rationality and perhaps other constraints, the interpreter *reaches* the best assignment of thoughts to their subject. Here, we might ask whether interpretationism takes the process to involve generating some large number of interpretations, and then whittling them down through further testing and the application of various desiderata. I have sometimes described interpretation as if this is involved. Depending on how many hypotheses have to be generated, this may be a process the ordinary interpreter could not complete, and therefore a significant idealisation. Alternatively, we might suppose that interpretationism should say that the process either must or could involve starting with an assumption of a very high degree of similarity between interpreter and subject, and then retreating from that assumption gradually if problems in interpretation arise. This possibility is suggested in Moran (1994), and suggests a process that may be more

feasible for real interpreters. Future investigation should consider this issue, and weigh its significance.

Two further avenues for future research on this issue could include comparisons with explanations of other skills and information gathering strategies, such as motor skills and ordinary ways of finding out about our environments, and an investigation of the general idea of a rational reconstruction.

The question of the precise relationship between the process invoked by interpretationism and the way we really find out about other people's thoughts has not been answered, then. However, there are features of interpretationism which suggest a close relationship between the two, and the discussions in this thesis go some way towards making an investigation of the question more tractable.

2. Knowledge of minds

Suppose that we can develop a good theory about the relationship between the interpretation in interpretationism and our everyday practices. We should then move on to how interpretationism affects our epistemology.

In chapter 1, section 3.1 I considered two sceptical worries: the traditional problem of other minds, which questions how we can know that others have minds and what they are thinking and the problem posed by eliminative materialism, which claimed that many of our central beliefs about the mind (both about general features of minds and about our own and others' minds) might turn out to be false, and that our psychological concepts might turn out not to refer to anything. I will take these in reverse order.

If being interpretable is sufficient for thinking (and being interpretable as thinking particular thoughts is sufficient for thinking those thoughts) then eliminativism doesn't get a foothold. It doesn't matter what is going on inside our heads, or what

neuroscience might find out about how our brains work, because the application of our psychological concepts, and the truth of our beliefs about thinking, don't depend upon things that we can't already find out about.[107]

The sort of interpretationism developed in this thesis also supplies a response to another of the Churchlands' criticisms of our everyday thinking about minds. The Churchlands complain that folk psychology is a stagnating research project which hasn't advanced its ability to help us predict and control for thousands of years.[108] We can reply that this claim is based on the mistaken view that our ideas about the mind have just the same purpose as a scientific theory.

Interpretationism also gets rid of some epistemological concerns we might have about our knowledge of other minds. In the first instance, it does not allow the worry that we might know all about a creature's history, see them behaving in just the ways that we would behave given certain thoughts, and yet, because we can't see inside to the states that cause the behaviour, or to certain epiphenomenal states, still be wrong about that creature being a thinker.

Moreover, since the sorts of behaviour a creature could produce and the interpretations that could be given as a result are the criteria for having certain mental states, it seems that our methods should often be good ways to find out what others are thinking. There is more to think about here, however, if we are to determine the reliability of our everyday beliefs about other people's minds. In particular, this discussion must depend upon the precise form of the answer to the question which dominated the previous section. Still, there are several things we *can* say as a result of the arguments of this thesis about the ways in which we might be ignorant or wrong about minds and thinkers.

---

[107] At least, this is so in normal cases. Special cases, such as those involving locked-in patients, show that there is a good sense in which we can't always find out what someone is thinking without some change in them which we ordinary interpreters can't bring about by ourselves. However, this concedes no ground to eliminativism.

[108] See Churchland (1989).

In chapter 3, I suggested that the interpretationist need not claim that interpretation gives real human interpreters any way of determining exactly when mental states occur. The thoughts about holism presented in chapter 4 suggest that we may not always know whether the information we have about a subject and their interactions with others and their environment suffices to allow us to determine all of their thoughts. Chapter 5 suggested that even when we know a lot about the history of an individual, there is still the possibility of deception. In the same chapter, we encountered a case where the ordinary human might not know they were in the presence of a thinker, because that thinker was paralysed. Chapter 6 also suggests that a thinker we're acquainted with could turn out to be very different from us in unexpected ways, for example because it does its thinking on Mars, or because someone else planned that it should think the thoughts it does.

Failures of general and self-knowledge have not been explicitly discussed in this thesis, but they must also be possible. For example, we could allow for some mistaken general beliefs about minds by allowing that the ordinary interpreter might not fully realise some of the possibilities for certain mental states, because when explicitly formulating beliefs about the mental they focus too much on the typical cases and not enough on the wide range of possibilities for interpretable behaviour. For example, we might initially suppose that beliefs have to be responsive to reasons, but after considering, for example, pathological cases, we might realise that our concept of belief was a bit more flexible.

None of these possibilities for ignorance and error seem at all unreasonable. Rather, interpretationism appears to make our beliefs about minds fallible in just the ways that we ordinarily suppose that they are fallible. As I pointed out in chapter 1, section 3.1, the thought that our beliefs about minds enjoy protection from refutation could be taken too far. Interpretationism does not do this. I suggest that it may offer just the right amount of protection for our beliefs about thought.

3. Avoidance of chauvinism

Finally, we must consider whether interpretationism avoids being a chauvinist theory, i.e. whether it avoids unfairly denying some being the status of thinker because of factors that ought to be irrelevant.

The worry in chapter 1 was that positions in the 'Yes' camp would deny that a being, like one of the Analogoids in the example of that chapter, could be a thinker, just because it had the wrong sort of physical stuff inside its head (or equivalent), despite what was claimed to be far more important evidence of thought, such as complex behaviour and successful interpersonal interactions. Interpretationism was designed to meet this problem, and it does: to use Dennett's phrase, it is 'maximally neutral about the internal structures that accomplish the rational competences it presupposes'. (2009: 346) Some would even say that it meets the problem a little too well, and lapses into liberalism (counting too many things as thinkers).[109]

However, there is another potential problem of chauvinism: a worry that we might unfairly deny a creature the status of thinker because it *thinks* rather differently from us. One might suppose that interpretationism makes this problem more pressing, rather than helping us to overcome it.

One source of this worry is the interpretationist's focus on rationality.[110] One might worry that this will lead the interpretationist to place too much emphasis on *our* standards for good thinking, and to unreasonably rule out the possibility of other thinkers who have substantially different concepts, principles, standards, intuitions and habits of thought from us. This worry is linked to the debate about the

[109] This was the problem discussed in chapter 7, where I argued that the interpretationist should accept the possibility of certain strange thinkers, and tried to mitigate or discredit the intuitions that some people have against them.

[110] That this might lead to chauvinism is a concern expressed in Stich (1990).

possibility of differing conceptual schemes, and a full discussion of the issue would need to address the nature and possibility of such schemes.[111]

Given the conclusions of this thesis, however, we can note reasons to be optimistic that interpretationism need not commit us to this second sort of chauvinism. This charge of chauvinism would look most plausible if we had adopted something like the Standard Picture of rationality, which saw being rational as a matter of following certain rules. However, I have suggested that the interpretationist adopt an account closer to that of Cherniak (1986). This demands that thinkers get at least some things right, but these need not be things that we find it at all easy to get right. It also demands that thinkers display some patterns in reasoning and acting, but allows that this might be quite different from the patterns we display. Such differences between us and other thinkers might make interpretation extremely difficult, but interpretationism need only claim that interpretation is possible, not that it is easy. The possibility is open that with the right interactions and probably quite a lot of imagination and creativity, we could come to understand subjects who think *very* differently from us.

This brings us to another source of the worry, however: the intuition that there are possible cases where there would be a very good sense in which we ordinary humans *couldn't* come to understand many of another thinker's thoughts because we lack, and could not through ordinary means gain, the cognitive resources necessary. An example might be the case of a thinker who was substantially more intelligent than us. There appears to be a problem with calling such a thinker interpretable without removing our theory too far from our limited human perspective.

This is a significant problem, but the discussion in this thesis suggests that it might also be tractable. In chapter 5 I argued that the sense in which an interpreter

[111] Davidson's thoughts on the subject are to be found in his 'On the Very Idea of a Conceptual Scheme', where he famously denies that alternative conceptual schemes are possible and therefore urges us to abandon the notion altogether. For an account of conceptual relativity which may be compatible with interpretationism, see Button (forthcoming), chapters 18 and 19.

'could' interpret a subject might be quite weak, and yet the subject still count as interpretable in the sense relevant to interpretationism. A similar strategy might be available here. For example, the interpretationist might say that the important thing is that interpretation be possible using the same basic method as that usable by the ordinary human interpreter: i.e. employing the assumption that a subject is at least minimally rational, and attempting to attribute thoughts through the attempt to understand the subject substantially in terms of their reasons using information about the subject's behaviour and environment. Thus we may be able to allow the possibility of thinkers too clever for the ordinary, or even the cleverest of humans to fully interpret by saying that our methods of interpretation are up to discovering their thoughts, although we are not up to employing those methods as they should be employed in some *importantly limited* range of cases.

Again, this potential response needs further development, but indicates at least that the possibility of thinkers who we humans are too stupid to fully interpret does not obviously show that interpretationism is a chauvinist theory of mind.

4. Conclusion

Evaluating the extent to which interpretationism exhibits the advantages claimed for a theory in the 'No' camp is not an easy task; and I have not offered a complete discussion of the issues involved. Nevertheless, the account developed through the course of this thesis is of use in answering such questions, and shows promise in each of the areas discussed. The discussion has also served to indicate important areas for future research.

There are many ways in which interpretationism needs further development. Given the advantages of the theory as developed so far, I suggest that further research on the subject would be worthwhile.

Appendix 1: Sitting on the Fence

In chapter 1, section 2 I suggested that there was a position which resisted classification in terms of the 'Yes' and 'No' camps I distinguished in answer to the question

> In order to count as a thinker, must a being possess particular sorts of inner states, mechanisms, events and/or structures?

This awkward position might count as a version of interpretationism, since it accepts the claim that reflecting on interpretation has an important role to play in elucidating thought. It does not, however, agree with the interpretationism I have characterised about what exactly about the mental interpretationism can help us to understand. Here, I outline what this theory says, and then explain why I do not consider it in any depth.

There could be an account according to which a creature is a thinker iff it can be identified as a thinker by an interpreter, and at least some of its thoughts can be identified by that interpreter. However, this account might say that there may nevertheless be some thoughts which an interpreter cannot identify. According to such a position, being a thinker would always involve having some interpretable thoughts, but not every thought would need to be interpretable in any sense of the word. The position I have in mind is at least similar to that suggested in Block (1981) as a behaviourist conception of intelligence.

Such a position must say something about why the partially interpretable creature counts as having the non-interpretable mental states that it does. Such an answer might refer in some way to the internal physical organisation of the creature. For example, it might suggest that the interpretable behaviour of a creature

determines what internal stuff is relevant to the creature's mental setup, but the arrangement of this internal stuff can then go on to determine the existence of thoughts which are not interpretable. On such a theory all thoughts would ultimately depend on the existence of interpretable thoughts, but what goes on inside the creature would also be important. For this reason, this position appears to straddle the fence between the 'Yes' and 'No' camps.

It is possible that an interesting position could be developed along these lines. However, I do not consider it further in this thesis for three reasons: first, because the problem of explaining why a creature counts as having certain non-interpretable thoughts seems substantial; second, because it does not seem to be a position espoused by any real interpretationists; and finally, because my stated aim is to consider what the most plausible view within the 'No' camp should look like.

I will call the theory suggested here *partial interpretationism*. Throughout the thesis, when I just use the word 'interpretationism', I refer to the view which I outline in chapter 3: the one which does fit more neatly into the 'No' camp.

Appendix 2: Table of Varieties of Interpretationism

Type	Derivative	Analytic	Dependence	Cartographical
Aim	To analyse thought using independent terms, and then to show that this has the consequence of making thought interpretable	To analyse thought in terms of interpretability	To analyse both thought and interpretation/ interpretability in independent terms	To illuminate the concept of thought, along with other concepts including interpretation, without giving an analysis of any one or arranging them in a hierarchy
Constitutive vs. Non-constitutive[112]	Non-constitutive	Constitutive	Constitutive	Constitutive?
Where rationality fits in	In the analysis of thought	As a consequence of the analysis of thought/ as part of our account of interpretation	In the joint analysis of thought and interpretability	It will appear in the account on the same level as thought and interpretation

[112] As these terms are defined by Child (1994).

Ambition as a form of interpretationism	Low	High	Moderate?	High
Problems	Giving the analysis	Giving the analysis, particularly without circularity	Showing that thought and interpretability can't be analysed separately; giving the analysis	Avoiding vicious circularity and achieving illumination.

Appendix 3: The Worth of the Cartographic Approach

As I have explained, cartographic interpretationism does not attempt to explain exactly what thought amounts to using independent terms. It can therefore use a notion like rationality to illuminate the concept of thought, and then go on to explain this notion of rationality by talking about particular kinds of thoughts or mental states or ideas, characterised as such. It might seem that this makes cartographic interpretationism a much easier project to pursue. It also opens the theory up to a very obvious criticism: that it will be circular, and therefore uninformative. If one does not understand the notion of thought already, then an explanation of it which invokes that very concept surely will not help you. Moreover, putting forward such an account can only distract us from the genuine explanations that are available: it is worse than useless.

A famous example of a circular explanation comes from Molière: he presents us with a doctor who explains why opium has a soporific effect by saying that it has a dormitive virtue in it whose nature is to cause the senses to become drowsy.[113] However, this is a causal explanation. A better parallel to cartographic interpretationism would be given by the following statement: What it means for something to have a soporific effect is for it to contain a dormitive virtue.

In the original example, we can at least understand the temptation for someone to give such a response if they don't know why opium is a soporific. And the claim does at least contain the assertion that there is a cause, and that it involves a property of opium. The revised version, on the other hand, just seems completely pointless. If cartographic interpretationism makes a claim like this, it too will be in

[113] From *Le Malade Imaginaire* (1879).

serious trouble. I will therefore consider what makes the revised statement so bad, and will argue that the cartographic interpretationist account of the mind does not share these flaws.

There are at least three reasons why the statement 'What it means for something to have a soporific effect is for it to contain a dormitive virtue' should count as viciously circular:

i) It introduces a new concept to define the one the questioner is confused about, and this new concept has no content beyond the one it's supposed to be defining.
ii) The statement has no consequences.
iii) There is a much more illuminating answer readily available. If someone speaks English, but doesn't know what it means for something to have a soporific effect, we can just tell them that it means the thing will send you to sleep.

Cartographic interpretationism fails to share these feature as follows:

i) If the cartographic interpretationist adopts the suggestions I have offered through this thesis so far, then he will try to illuminate what it is to have a thought or be a thinker by calling on a lot of different concepts. He will call on rationality, getting things right, responsiveness, reasons, points of view, patterns in behaviour, patterns in reasoning, inference, consistency, goals, achievements, disagreement, excuse, interpretation, interpretability, and sincerity, to name a few. These are not new concepts; they are ones which we use often, and usually quite successfully[114], even though it may be difficult to state precise rules for how they should be used. Moreover, these concepts do not all just amount to the same thing, and we use them in quite different ways. Cartographic interpretationism will aim to show how they

[114] A possible exception to this is the notion of interpretability. However, Part II aimed to define this in terms of other concepts which are more familiar.

relate to one another, and thus to provide a better understanding of many of them, without identifying them with each other.

Thus, if we want to say that the cartographic interpretationist's answer to the question 'What does it take to be a thinker?' is circular, we should note that it does at least provide us with a very big circle. We should also note that, although the project will not face exactly the same problems as an analysis, and we will have different tools to deal with the problems we do face, nevertheless it is clearly a difficult job to map out relationships between so many substantial concepts. Particular proposals will face objections other than circularity, and the cartographic interpretationist will have to meet them.

ii) The theory does have consequences. Two in particular have been prominent in this thesis: the anti-chauvinism of interpretationism, leading us to count some perhaps quite unexpected things as thinkers (cf. chapter 6); and the protection the theory offers us against eliminativism, with a corresponding suggestion for how to view the relationship between interpersonal interpretation and scientific investigation of the mind, and thus also between the philosophy and the science of the mind.

iii) Finally, the question 'What does it take to be a thinker?' does not have another easy and obvious answer, as 'What makes something soporific?' did. There are indeed other potential answers: there are the answers provided by the positions in the 'Yes' camp, and there are interpretationist analyses like the one suggested in section 1. Cartographic interpretationism will only provide the best and fullest available answer to the question if there is something wrong with these other projects. However, it is far from clear that the other projects will work, and so we have no reason to conclude against cartographic interpretationist on this basis just yet.

This short discussion cannot give a complete account of a way of engaging in the philosophy of mind. However, if you are a person who thinks that there can be informative philosophical projects which do not count as analyses, and philosophical accounts which provide us with virtuous and informative circles, then I commend cartographic interpretationism as a candidate for just such a project.

REFERENCES

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

Bennett, J. (1964) *Rationality: An Essay Towards an Analysis,* London: Routledge and Keegan Paul Ltd.

Bennett, J. (1976) *Linguistic Behaviour*, Cambridge: Cambridge University Press

Ben-Yami, H. (2005) 'Behaviourism and Psychologism: Why Block's Argument Against Behaviourism is Unsound' *Philosophical Psychology*, 18 (2): 179-186

Block, N. (1980a) 'What intuitions about homunculi don't show', *Behavioural and Brain Sciences*, 3 (3): 425-426

Block, N. (1980b) *Readings in Philosophy of Psychology*, Cambridge: Harvard University Press

Block, N. (1981) 'Psychologism and Behaviourism' in *Philosophical Review* 90 (1): 5-43

Boden, M. (1994) *Dimensions of Creativity*, Cambridge, MA: MIT Press

Bortolotti, L. (2005) 'Intentionality Without Rationality' in *The Proceedings of the Aristotelian Society*, New series, 105: 369-376

Braddon-Mitchell, D. and Jackson, F. (2007) *Philosophy of Mind and Cognition*, second edition, London: Blackwell

Button, T. (forthcoming) *The Limits of Realism*, Oxford: Oxford University Press

Byrne, A. (1998) 'Interpretivism' in *European Review of Philosophy*, 3: 199-223

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

Child, W. (1993) 'Anomalism, Uncodifiability, and Psychophysical Relations' in *The Philosophical Review*, 102 (2): 215-245

Child, W. (1994) *Causality, Interpretation and the Mind*, Oxford: Clarendon Press

Child, W. (2007) 'Dreaming, Calculating, Thinking: Wittgenstein and Anti-realism about the Past' in *The Philosophical Quarterly*, 57 (227): 252-272

Churchland (1981) 'Eliminative Materialism and Propositional Attitudes', *Journal of Philosophy*, 78 (2): 67-90

Churchland, P.M. (1989) 'Folk Psychology and the Explanation of Human Behaviour', in *Philosophical Perspectives*, 3: 225-241

Davidson, D. (1980) *Essays on Actions and Events*, Oxford: Clarendon Press

Davidson, D. (1984). *Inquiries into Truth and Interpretation*, Oxford: Clarendon Press

Davidson, D. (2001). *Subjective, Intersubjective, Objective*, Oxford: Clarendon Press

Davidson, D. (2004) *Problems of Rationality*, Oxford: Clarendon Press

Davidson, D. (2005) *Truth, Language and History*, Oxford: Clarendon Press

Davies, M. (1991) 'Concepts, Connectionism, and the Language of Thought' in Ramsey, Stich and Rumelhart ed. *Philosophy and Connectionist Theory*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Dennett, D. C. (1971) 'Intentional Systems,' in *Journal of Philosophy*, 68 (4): 87-106

Dennett, D.C. (1976) 'Are Dreams Experiences?' in *The Philosophical Review*, Vol. 58, No. 2: 151-171

Dennett, D.C. (1978) *Brainstorms: Philosophical Essays on Mind and Psychology*, Montgomery, VT: Bradford Books

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

Dennett, D.C. (1998) *Brainchildren: essays on designing minds*, London: Penguin

Dennett, D.C. (2009) 'Intentional Systems Theory' in McLaughlin, Beckerman and Water ed. Oxford Handbook of the Philosophy of Mind, Oxford: Oxford University Press

Ellis, J. (2011) 'The Relevance of Radical Interpretation to the Understanding of Mind' in J. Malpas ed. *Dialogues with Davidson: Acting, Interpreting, Understanding*, Cambridge, MA: MIT Press

Fodor, J. (1976) *The Language of Thought*, Cambridge, MA: Harvard University Press

Fodor, J. (1987) *Psychosemantics*, Cambridge, MA: MIT press

Gardner, S. (1993) *Irrationality and the Philosophy of Psychoanalysis*, Cambridge: Cambridge University Press

Gerrans, P. (2004) 'Cognitive Architecture and the Limits of Interpretationism' in *Philosophy, Psychiatry and Psychology*, 11 (1): 43-48

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

Gregory, J.C. (1916) 'Dreams as Psychical Explosions' in *Mind*, New Series, 25 (98): 193-205

Heal, J. (1994) 'Semantic Holism: still a Good Buy' in *Proceedings of the Aristotelian Society*, 94: 325-339

Heal, J. (2003) *Mind, Reason and Imagination: selected essays in philosophy of mind and language*, Cambridge: Cambridge University Press

Heal, J. (2007) 'Back to the Rough Ground! Wittgensteinian Reflections on Rationality and Reason', *Ratio* 20 (4):403–421

Hobson, J. A. (2005) *Dreaming: A Very Short Introduction*, Oxford: Oxford University Press

Hooker, C. A. (1994) 'Idealisation, Naturalism, and Rationality: Some Lessons from Minimal Rationality' in *Synthese*, 99 (2): 181-231

Hume, D. (2007) *An Enquiry Concerning Human Understanding And Other Writings*, Cambridge: Cambridge University Press

Jackman, H. (1999) 'Moderate Holism and the Instability Thesis' in *American Philosophical Quarterly*, 36 (4): 361-369

Jackson, F. and Pettit, P. (1993) 'Folk Belief and Commonplace Belief' in *Mind and Language*, 8 (2): 298-305

Kacelnik, A. (2006) 'Meanings of rationality' in Hurley and Nudds ed. *Rational Animals*, Oxford: Oxford University Press

Kahneman, D. and Tversky, A. (1983) 'Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgement,' in *Psychological Review*, 90 (4): 293-315

Kahneman D. (2011) *Thinking Fast and Slow*, London: Penguin

Knobe, J. (2003) 'Intentional action in folk psychology: An experimental investigation' in *Philosophical Psychology*, 16 (2): 309-324

Knobe, J. (2006) 'The Concept of Intentional Action: A Case Study in the Uses of Folk Psychology' in *Philosophical Studies*, 130 (2): 203-231

Kolodny, N. (2005) 'Why Be Rational?' in *Mind*, New Series, 114 (455): 509-563

Lehrer, J. (2012) 'Is Self Knowledge Overrated?' in *The New Yorker*, Condé Nast, available at http://www.newyorker.com/online/blogs/books/2011/10/is-self-knowledge-overrated.html#ixzz213l0APzd

Lepore, E. and Ludwig, K. (2005) *Donald Davidson: Meaning, Truth, Language and Reality*, Oxford: Clarendon Press

Lewis (1966) 'An Argument for the Identity Theory' in the *Journal of Philosophy*, 63 (1):17-25

Lewis, D. (1974) 'Radical Interpretation' in his 1983 ed. *Philosophical Papers, Volume 1*, New York: Oxford University Press

Malcolm, N., (1957) 'Dreaming and Scepticism: A Rejoinder', *Australasian Journal of Philosophy*, 35 (3): 207–211

Malcolm, N. (1959) *Dreaming, London:* Routledge & Kegan Paul Ltd.

Molière, J. P. (1879) *Le Malade Imaginaire*, translated by Roscoe Mongan, London: James Cornish and Sons

Moran, R. (1994) 'Interpretation Theory and the First Person' in *the Philosophical Quarterly*, 44 (175): 154-173

Morton, A. (2003) *The Importance of Being Understood: Folk Psychology as Ethics*, London: Routledge

Mullane, H. (1983) 'Defence, Dreams and Rationality' in *Synthese*, 57 (2): 187-204

Orenstein, D. (2011) 'BrainGate neural interface reaches 1,000-day milestone' available at http://news.brown.edu/pressreleases/2011/03/braingate,

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

Peacocke, C. (1983) *Sense and Content*, Oxford: Clarendon Press

Peacocke, C. (1997) 'Holism' in B Hale and C. Wright ed. *A Companion to the philosophy of Language*, Oxford: Blackwell

Putnam, H. (1975) *Philosophical Papers*, Cambridge: Cambridge University Press

Quine, W. (1960) *Word and Object*, Cambridge, MA: MIT Press

Rescorla, M. (forthcoming) 'Rationality' in Lepore, E. and Ludwig, K. ed. *A Companion to Davidson*, Oxford: Wiley-Blackwell

Ryle, G. (1949) *The Concept of Mind*, London: Hutchinson

Samuels, R., Stich, S., and Bishop, M. (2002) 'Ending the rationality Wars: How to make Disputes About Human Rationality Disappear' in Elio ed. *Common Sense, Reasoning and Rationality*, New York: Oxford University Press

Samuels, R., Stich, S. and Faucher, L. (2004) 'Reason and Rationality' in Niiniluoto, I., Sintonen, M. and Wolenski, J., ed. *Handbook of Epistemology*, Doerdrecht: Kluwer

Schnakers, C., Perrin, F., Schabus, M., Hustinx, R., Majerus, S., Moonen. G., Boly, M., Vanhaudenhuyse, A., Bruno, M. A., Laureys, S. (2009) 'Detecting consciousness in a total locked-in syndrome: an active event-related paradigm' in *Neurocase* 15 (4): 271-7

Schroeder, S. (1997) 'The Concept of Dreaming: On Three Theses by Malcolm' in *Philosophical Investigations* 20 (1): 15-38

Searle, J. (1980) 'Minds, Brains and Programs', *Behavioural and Brain Sciences*, 3 (3): 417–457

Searle, J. R. (1992) *The Rediscovery of the Mind*, Cambridge MA: MIT Press

Sosa, E. (2005) 'Dreams and Philosophy' in *Proceedings and Addresses of the American Philosophical Association*, 79 (2): 7-18

Stanovich, K. E. and West, R. F. (2000) 'Individual differences in reasoning: Implications for the rationality debate' in *Behavioural and Brain Sciences*, 23 (5): 645–665

Stein, E. (1994) *Without Good Reason*. Oxford: Clarendon Press

Stich, S. (1990) *The Fragmentation of Reason*, Cambridge, MA: MIT Press

Turing (1950) 'Computing machinery and intelligence' in *Mind*, 59 (236): 433-460

Verheggen, C. (1997) 'Davidson's Second Person', *The Philosophical Quarterly*, 47 (188): 361-369

Walton, K. (1990) *Mimesis as Make-Believe: on the foundations of the representational arts*, Cambridge, MA: Harvard University Press

Wittgenstein, L. (1953) *Philosophical Investigations*, Oxford: Blackwell

Wittgenstein, L. (1981) *Zettel*, Oxford: Blackwell